nature astronomy

Article

https://doi.org/10.1038/s41550-025-02579-7

Incorporation of model accuracy in gravitational wave Bayesian inference

Received: 1 October 2024

Accepted: 12 May 2025

Published online: 15 July 2025

Check for updates

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

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 new method for incorporating the accuracy of several models into 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 further apply our method to a real gravitational wave signal and, when assuming the binary black hole hypothesis, demonstrated that the source of GW191109_010717 has unequal component masses, with a 69% probability for the primary being above the maximum black hole mass from stellar collapse. We envisage that this method will become an essential tool for ground-based gravitational wave astronomy.

Our ability to infer the properties of colliding black holes from an observed gravitational wave (GW) signal depends 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 can be obtained only with models that are perfect descriptions of general relativity (assuming a known understanding of the noise in the GW detectors[11–13]).

Unfortunately, directly computing GW signals from general relativity is a computationally expensive task; numerical relativity simulations, for which Einstein's equations of general relativity are solved on high-performance computing clusters, require millions of central processing unit (CPU) hours to perform[14]. For this reason, only several thousand simulations are currently available[14–21]. As a result, the latest GW models rely on analytical or semi-analytical prescriptions that are calibrated to the numerical relativity simulations[2,3,22–26] or are based on surrogate modelling techniques[27,28]. However, each modelling approach will incur some degree of approximation errors.

The accuracy of a GW model is typically measured by the mismatch[29] 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 the 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 in which results from numerous analyses are combined. A Bayesian analysis is performed for each GW model and the results are

[1]Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth, UK. [2]School of Mathematics & Statistics, University College Dublin, Dublin, Ireland. [3]Theoretical Astrophysics Group, California Institute of Technology, Pasadena, CA, USA. [4]Mathematical Sciences & STAG Research Centre, University of Southampton, Southampton, UK. ✉e-mail: charlie.hoy@port.ac.uk

https://doi.org/10.1038/s41550-025-02579-7

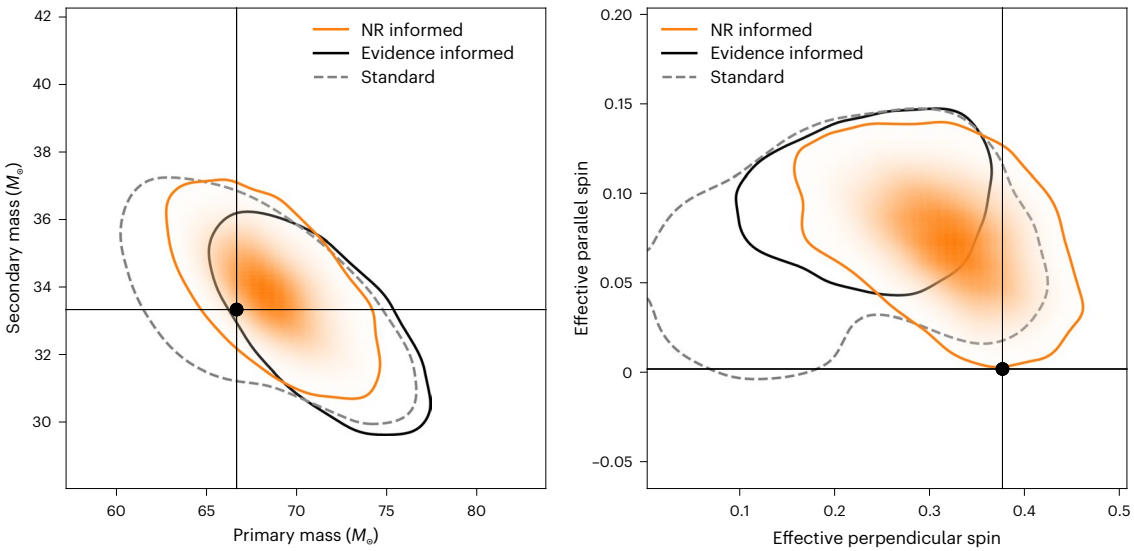

Fig. 1 | Two-dimensional posterior probabilities obtained in our analysis of the SXS:BBH:0926 numerical relativity simulation. Left, measurement of the primary and secondary masses of the binary. Right, inferred effective parallel and perpendicular spin components (as defined in Methods; see equations (11) and (12)). An effective perpendicular spin of 0 means that the spin vector lies perpendicular to the plane of the binary. The contours represent 90% credible intervals, and the black cross hairs indicate the true values.

either mixed together with equal weights[30] or according to their relative Bayesian evidence[31] or by averaging the likelihood[32]. An alternative technique involves sampling over a set of GW models in a single joint Bayesian analysis[33,34]. 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[35–39]. Either these methods have not been demonstrated in practice or they are suitable only for a single model. Preliminary work has investigated incorporating model accuracy into likelihood averaging techniques for simplified models[40]. However, this approach incurs a comparable computational cost to evidence mixing[31] and is difficult to interface with standard Bayesian inference techniques.

In this work we present an approach for incorporating the accuracy of several cutting-edge models into a single GW Bayesian analysis while also reducing the computational cost (Methods). This technique accounts for modelling errors by prioritizing the most accurate GW model in each region of parameter space, thereby mitigating against biased results from using models that are unfaithful to general relativity. For GW signals likely observed by the LIGO[11]–Virgo[12]–KAGRA[13] GW detectors, we demonstrate that current techniques will more likely 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.

GW Bayesian inference

We first applied our approach to analyse an example of a theoretical GW signal expected from general relativity, specifically, the SXS:BBH:0926 numerical relativity simulation[16,41] produced by the Simulating eXtreme Spacetimes Collaboration (https://www.black-holes.org). We assumed a total mass of 100 $M_\odot$, and we injected 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 dimensionless spin magnitudes perpendicular to the orbital angular momentum (within the orbital plane of the binary) for both black holes of ~0.8 out of a maximum possible value of 1. For this system, the general relativistic phenomenon of spin-induced orbital precession[42] is substantial and contributes a signal-to-noise ratio[43] of ~9 to the total power of the signal. This simulation was chosen because most GW models obtain biased results and disagree on the inferred binary parameters[5,7]. Such a system with notable spin-induced orbital precession has been predicted to be observed once in every 50 GW observations made by the LIGO, Virgo and KAGRA GW observatories based on current black hole population estimates[44].

We used three of the most accurate cutting-edge models currently available for describing the theoretical GW signals produced by colliding black holes: IMRPhenomXPHM[22] (with the updated precession formulation[26]), IMRPhenomTPHM[23] and SEOBNRv5PHM[2]. All models include the general relativistic phenomenon of spin-induced orbital precession[42] and higher-order multipole moments[45]. We analysed 8 s of data and considered frequencies only in [20, 2,048] Hz. We generated the numerical relativity simulation from ~10 Hz to ensure that most of the higher multipole content was generated before our analysis window. Our analysis was restricted to a two-detector network comprising LIGOHanford and LIGOLivingston[11], and we assumed a theoretical power spectral density for the design sensitivity of Advanced LIGO[46]. We used the most agnostic priors available for all parameters, which are identical to those used in all detections made by the LIGO–Virgo–KAGRA Collaboration[30]: flat in the component masses, spin magnitudes and cosine of the spin tilt angles. We performed Bayesian inference with the DYNESTY nested sampling software[47] using Bilby (ref. 48), as has been done in all LIGO–Virgo–KAGRA analyses since the third GW catalogue[30].

Figure 1 compares the results obtained with our method to those from two widely adopted techniques. The contours labelled NR informed (informed by numerical relativity) use the method presented here, evidence informed combines separate inference analyses obtained with different GW models according to their relative Bayesian evidence[31] and standard combines the results of a separate inference analysis with equal weights. Standard is the method currently adopted by the LIGO–Virgo–KAGRA Collaboration[30], as it is probably the most agnostic. When considering the inferred primary and secondary masses of the binary, all three techniques captured the true value within the two-dimensional marginalized 90% credible interval. Both the NR-informed and standard methods more accurately inferred 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 was inflated in comparison to the method presented here and to the evidence-informed result.

https://doi.org/10.1038/s41550-025-02579-7

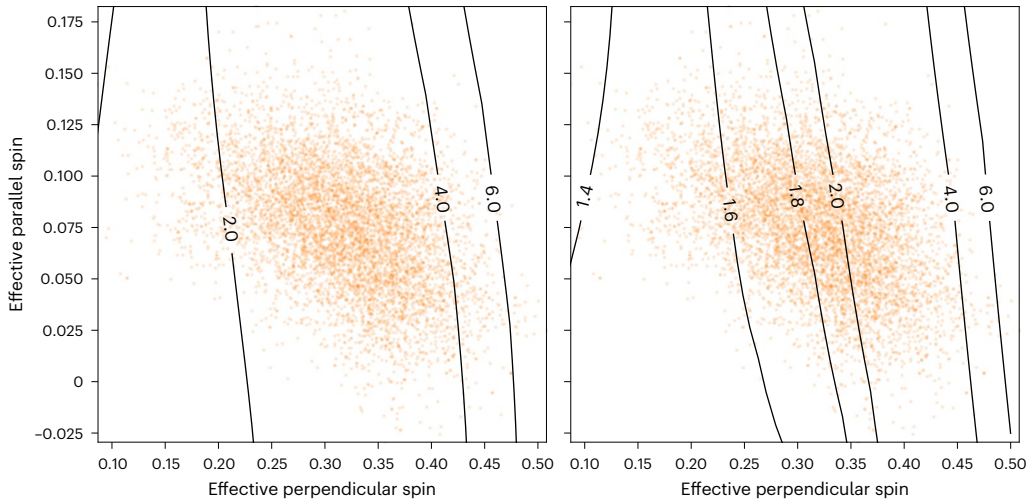

Fig. 2 | Ratio of mismatches to numerical relativity simulations. Contour plots showing the ratio of mismatches to numerical relativity simulations for different effective parallel and perpendicular spin components when averaging over different mass configurations. Left, comparison of IMRPhenomXPHM and SEOBNRv5PHM. Right, comparison of 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.

We now turn our attention to the inferred spin of the binary. Because the individual spin components are difficult to measure for binary black holes at present-day detector sensitivities[49], we considered the measurement of effective spin parameters that describe the dominant spin effects of the observed GW signal[50,51]. Figure 1 shows the measurement of the effective spin parallel and perpendicular to the orbital angular momentum, as defined in Methods. We see substantial 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[31], it produced an inaccurate result for this simulated signal. 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[34]. Our analysis, on the other hand, predominantly uses SEOBNRv5PHM, specifically 90% of the time, whereas IMRPhenomTPHM is used 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 region of the parameter space, there is no guarantee that this model is accurate enough to avoid biases in the inferred parameter estimates[52,53]. However, we highlight that it is the most accurate method of those currently used, and it can be evolved to include more accurate models when they are developed.

Figure 2 presents the ratio of mismatches obtained with the different GW models used in this work. SEOBNRv5PHM has the smallest mismatch in the region of parameter space containing the simulation parameters and is, therefore, the most faithful to the general relativity results in this region. Specifically, it yields mismatches ~3 and ~1.8 times smaller than IMRPhenomXPHM and IMRPhenomTPHM do, respectively.

Because the NR-informed approach 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, there is a notable decrease in the computational cost. Our analysis uses 30% fewer computational resources than the standard and evidence-informed analyses during sampling. 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-informed analyses.

Our technique is free to use any combination of GW models. When SEOBNRv5PHM was removed from this analysis, we found consistent results between our method and the evidence-informed result, 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 the parameter space.

A single analysis with the model that is, on average, the most accurate in the parameter space of interest can be performed[54]. However, 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 found that IMRPhenomTPHM is more accurate than SEOBNRv5PHM and that IMRPhenomXPHM is of comparable accuracy to SEOBNRv5PHM. By simply averaging the mismatch across the parameter space, we neglected this information, resulting in the use of a less accurate model in certain regions of the parameter space. On the other hand, the method presented in this work to incorporate the accuracy of several models into a single GW Bayesian analysis fully uses this information.

Numerical relativity surrogate techniques provide accurate models for describing GWs produced by colliding black holes[27,28]. We did not sample over surrogate models in this work because they are used as a proxy for numerical relativity simulations when assessing model accuracy (Methods). We quantified the efficacy of our approach by comparing its results to those obtained with surrogate models. For the same numerical relativity simulation we found that NRSur7dq4—the leading generic-spin numerical relativity surrogate model[27]—more accurately captures the true parameters of the binary, as expected (Supplementary Fig. 1). Our NR-informed approach offers the most statistically similar one-dimensional posterior probability distributions to the surrogate posteriors out of the methods considered in this work.

Contrary to standing belief, NRSur7dq4 is not guaranteed to be the most accurate model, even within its calibration region. For instance, when comparing against numerical relativity simulations that were not used to validate NRSur7dq4, we found that SEOBNRv5PHM, IMRPhenomTPHM and IMRPhenomXPHM can more faithfully

https://doi.org/10.1038/s41550-025-02579-7

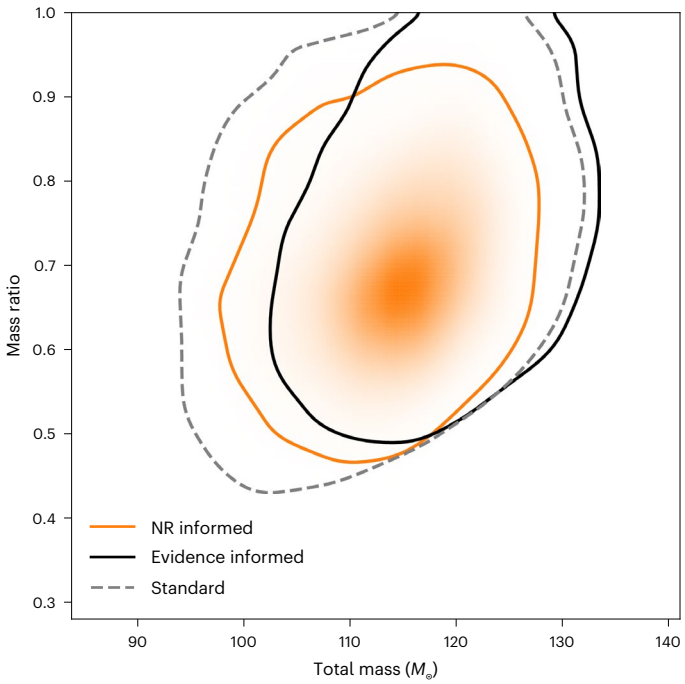

Fig. 3 | Two-dimensional posterior probabilities obtained in our analysis of GW191109_010717. Measured total mass of the binary and the mass ratio, defined as the secondary mass divided by the primary mass. A mass ratio of 1 implies equal binary component masses. The mass ratio is always less than or equal to 1. The contours represent 90% credible intervals.

describe numerical relativity than NRSur7dq4. Specifically, based on mismatches against the CF_52 simulation[14] (a single-spin, mass ratio 1:4 simulation with primary dimensionless spin magnitude 0.6 and total masses 75, 80 and 85 $M_\odot$), we estimated that according to our NR-informed approach, SEOBNRv5PHM would be ~140 times more likely to be used than NRSur7dq4 in this region of the parameter space due to its improved accuracy. Although corner cases such as this exist, NRSur7dq4 is still suitable as a proxy for numerical relativity in this work because we are performing Bayesian inference on numerical relativity simulations where the surrogate is the most accurate model[27]. We emphasize that when more numerical relativity simulations become available, the surrogate will no longer be needed as a proxy to assess model accuracy, and we will be able to incorporate the faithfulness of all models, including the surrogate, within our Bayesian framework.

Although not presented in this section (Supplementary Figs. 2 and 3), we also analysed the SXS:BBH:0143 numerical relativity simulation[16,41] and the SXS:BBH:1156 numerical relativity simulation[16,41] produced by the Simulating eXtreme Spacetimes Collaboration. SXS:BBH:0143 was chosen because 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 because it has largely asymmetric mass components and lies in the extrapolation regime of our technique (see Methods for details). We found for SXS:BBH:0143 largely overlapping posteriors between all three methods, with most of the one-dimensional marginalized 90% confidence intervals containing the true value. Our analysis of this case used SEOBNRv5PHM 80% of the time, IMRPhenomTPHM 15% of the time and IMRPhenomXPHM 5% of the time. This represents the worst-case scenario: by construction our method should at worst give the same results as other methods. For SXS:BBH:1156, our method outperformed the standard and evidence-informed analyses despite partly being in the extrapolation regime of our technique: we more accurately captured the true parameters of the binary. Like SXS:BBH:0926, the evidence-informed analysis preferred IMRPhenomTPHM owing to the larger Bayesian evidence, whereas our analysis preferred SEOBNRv5PHM, as it is the more accurate model in this region of the parameter space. Our analysis used SEOBNRv5PHM for 78% of the time, IMRPhenomTPHM for 9% and IMRPhenomXPHM for 13%.

Finally, we applied our technique to a real GW signal. GW191109_010717 was observed on 9 November 2019[30] and has sparked interest within the community because its source probably has large component masses that lie within the upper mass gap. Theories indicate that the maximum black hole mass from stellar collapse is ~65 $M_\odot$ (ref. 55). As shown in Fig. 3, by incorporating model accuracy in the GW Bayesian inference, we more tightly constrained the total mass of GW191109_010717 to 100 $M_\odot < M < 124\ M_\odot$ and demonstrated that the source of GW191109_010717 has conclusively unequal component masses (assuming the binary black hole hypothesis). Our reanalysis shows that when using consistent priors and sampler settings as the LIGO–Virgo–KAGRA Collaboration, there is a 69% probability that the primary component mass of GW191109_010717 lies within the upper mass gap, consistent with previous work where GW191109_010717 was reanalysed with NRSur7dq4 (ref. 56). Compared to the 51% probability from the LIGO–Virgo–KAGRA analysis[30], we have appreciably increased the probability that GW191109_010717 was produced from a hierarchical formation mechanism in which the primary component mass was formed from a previous black hole merger. Other one-dimensional posterior probability distributions remain comparable among the different methods considered in this work.

Conclusions

In this work we present a method for incorporating model uncertainty into GW Bayesian inference. We applied this method to theoretical GW signals expected from general relativity and show that (1) it marginalizes over model uncertainty by prioritizing the most accurate model in each region of the parameter space and that (2) it 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. Although the approach preferentially uses the most accurate model in each region of the parameter space, there is no guarantee that that model is accurate enough to avoid biases in the parameter estimates. However, GW models are continually being developed and will probably improve in accuracy across the parameter space. Once available, these more accurate models can be incorporated into this method. Similarly, when more numerical relativity simulations are produced, the accuracy of this method will increase and more models can be included. The method presented here is applicable to ground-based GW parameter estimation analyses, and we highly encourage its use.

Methods

Estimating waveform accuracy

As discussed in 'Model systematics in gravitational wave astronomy', 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 (h_m) of a signal and the signal itself (h_s)[29]:

$$\langle h_\mathrm{m} \mid h_\mathrm{s}\rangle = 4\Re \int_{f_\mathrm{min}}^{f_\mathrm{max}} \frac{\tilde{h}_\mathrm{m}^{*}\tilde{h}_\mathrm{s}}{S_n(f)}\,\mathrm{d}f, \tag{1}$$

where a tilde denotes a Fourier transform, an asterisk denotes complex conjugation and $S_\mathrm{n}(f)$ is the noise power spectral density, which in this work is the design sensitivity of Advanced LIGO[46]. The mismatch[29] between two signals is computed by optimizing the normalized inner product over a set of (intrinsic or extrinsic) model parameters $\boldsymbol{\lambda}_\mathrm{m}$:

$$\mathcal{M} = 1 - \max_{\boldsymbol{\lambda}_\mathrm{m}} \frac{\langle h_\mathrm{m} \mid h_\mathrm{s}\rangle}{\sqrt{\langle h_\mathrm{m} \mid h_\mathrm{m}\rangle\langle h_\mathrm{s} \mid h_\mathrm{s}\rangle}}. \tag{2}$$

The intrinsic parameter space for a generic quasi-circular compact binary system comprises two masses, $m_{1,2}$, and two spin vectors, $\mathbf{S}_{1,2}$, adding up to eight degrees of freedom (d.f.). Additionally, a quasi-circular binary comes with seven more extrinsic parameters: the right ascension, declination and the luminosity distance $\{\alpha, \delta, d_L\}$ to the centre of mass of the binary; the inclination of the orbit and its relative polarization $\{\iota, \psi\}$; and the overall constant time and phase shift $\{t_c, \varphi_c\}$ of the GW.

For binaries where the spins of the compact bodies are aligned with the orbital angular momentum of the system, 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 d.f.: $\{m_1, m_2, \mathbf{S}_{1z}, \mathbf{S}_{2z}\}$, 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 held all extrinsic parameters fixed and optimized the match over the set of model parameters $\boldsymbol{\lambda}_m = \{t_c, \varphi_c\}$. We maximized over t_c with an inverse fast Fourier transform and over ϕ_c using the Nelder–Mead optimization algorithm in the minimize function in SciPy[57].

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 was to isolate the intrinsic parameter space, so the mismatch to which we intended to fit had somehow to be independent of the extrinsic parameters. For this purpose, we first mapped $\{\alpha, \delta, \psi\}$ into a single parameter known as the effective polarizability κ (ref. 58). We then prepared an evenly spaced signal grid over the $\{\kappa, \phi_c, \iota\}_s \in [0,\pi/2] \otimes [0, 2\pi) \otimes [0, \pi]$ space with $7 \times 6 \times 7 = 294$ elements. At each point in this signal grid, we computed the sky-optimized mismatch[5,58,59] between the signal and the model template from equation (2), where the parameter set we optimized over is $\boldsymbol{\lambda}_m = \{t_c, \varphi_c, \kappa, \varphi_{spin}\}$. Here φ_{spin} represents the freedom to rotate the in-plane spin (azimuthal) angles ϕ_1 and ϕ_2 of $\mathbf{S}_1$ and $\mathbf{S}_2$ by a constant amount. κ was optimized analytically, and φ_c and φ_{spin} were optimized numerically using dual annealing algorithms[3,22,59]. Note that there is no universally agreed grid for $\{\kappa, \varphi_c\}$[26,59,60] nor for ι (refs. 3,61). Our specific choice for the $\{\kappa, \varphi_c\}$ grid was based on recent work[5]. Our ι grid spacing is also consistent with results in the literature[61] but extended to π because the 'up/down' symmetry of the GW multipoles with respect to the orbital plane is broken due to precession[62–67].

With these optimizations, we arrived at the maximum possible match between the template and the signal at a given point $\{\kappa, \phi_c, \iota\}_s$ in the signal grid. We repeated this procedure at every point of the 294-element grid and then computed the mean of this set as our final result for the mismatch:

$$\mathcal{M}_{av} := \frac{1}{294} \sum_{s=1}^{294} \mathcal{M}(\kappa_s, \varphi_{c,s}, \iota_s). \tag{3}$$

This was done to marginalize over any dependence of the mismatch on the sky position and inclination, thus obtaining values that depend exclusively on the intrinsic parameters of the source. We additionally retained the standard deviation σ of the 294-mismatch set and used this as our error bar when needed. Note that our mean match, $1 - \mathcal{M}_{av}$, is a discretely averaged version of the sky-and-polarization-averaged faithfulness given by equation 35 of ref. 2. For the remainder of this article, we drop the subscript 'av' from $\mathcal{M}$.

Multi-model Bayesian inference

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

$$p(\boldsymbol{\lambda} \mid d, \mathfrak{M}_i) = \frac{\Pi(\boldsymbol{\lambda} \mid \mathfrak{M}_i)\, \mathcal{L}(d \mid \boldsymbol{\lambda}, \mathfrak{M}_i)}{\mathcal{Z}}, \tag{4}$$

where $\Pi(\boldsymbol{\lambda} \mid \mathfrak{M}_i)$ is the probability of the parameters $\boldsymbol{\lambda}$ given the model $\mathfrak{M}_i$, otherwise known as the prior; $\mathcal{L}(d \mid \boldsymbol{\lambda}, \mathfrak{M}_i)$ 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, $\mathcal{Z} = \int \Pi(\boldsymbol{\lambda} \mid \mathfrak{M}_i)\, \mathcal{L}(d \mid \boldsymbol{\lambda}, \mathfrak{M}_i)\, \mathrm{d}\boldsymbol{\lambda}$, otherwise known as the evidence. It is often not possible to trivially evaluate the model-dependent posterior distribution; the challenge is evaluating the evidence because that involves computing the likelihood times the 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[68]. 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 likelihood-constrained prior; the new point is randomly drawn from the prior provided that the likelihood is larger than the point that it is replacing. This iterative process continues until the highest likelihood region(s) is identified.

When there is an ensemble of models, Bayesian model averaging can be used to marginalize over the model uncertainty:

$$\begin{aligned} p(\boldsymbol{\lambda} \mid d) &= \sum_{i=1}^{N} p(\boldsymbol{\lambda} \mid d, \mathfrak{M}_i)\, p(\mathfrak{M}_i \mid d) \\ &= \sum_{i=1}^{N} \left[\frac{\mathcal{Z}_i\, \Pi(\mathfrak{M}_i)\, p(\boldsymbol{\lambda} \mid d, \mathfrak{M}_i)}{\sum_{j=1}^{N} \mathcal{Z}_j\, \Pi(\mathfrak{M}_j)} \right], \end{aligned} \tag{5}$$

where $p(\mathfrak{M}_i \mid d)$ is the probability of the model $\mathfrak{M}_i$ given the data, $\Pi(\mathfrak{M}_i)$ is the discrete prior probability for the choice of model and N is the number of models in the ensemble. If there are uniform priors for the model, $\Pi(\mathfrak{M}_i) = 1/N$, Bayesian model averaging simply averages the model-dependent posterior distributions, weighted by the evidence.

An alternative to marginalizing over model uncertainty is to simultaneously infer the model and model properties in a single joint analysis[34]. Here, the parameter set $\boldsymbol{\lambda}$ is expanded to include the model m: $\tilde{\boldsymbol{\lambda}} = \{\lambda_1, \lambda_2, \ldots, \lambda_j, m\}$, and a discrete set of models can be sampled during standard Bayesian inference analyses: for each step in, for example, a nested sampling algorithm, a $(j+1)$-dimensional vector of model parameters is drawn from the prior, including an integer for the model, m. The integer m is mapped to a GW model, and the likelihood is evaluated by passing the remaining model parameters and the selected model to the standard GW likelihood[1]. It has been demonstrated that such a joint analysis will be at most N times faster to compute compared to performing Bayesian model averaging[34].

Defining a discrete prior probability for a model in GW astronomy is challenging because the accuracy of each model varies across the parameter space $\boldsymbol{\lambda}$ (ref. 5). This makes it difficult to perform Bayesian model averaging; a uniform prior probability is often assumed for the choice of model[31,32] or, in some cases, the model accuracy is averaged over the parameter space of interest[34]. However, a parameter-space-dependent prior for the choice of model may solve this problem[34]. For instance, a j-dimensional vector of model parameters can be drawn from the prior and $\Pi(\mathfrak{M}_i \mid \boldsymbol{\lambda})$ can be evaluated for all models, that is, the prior probability of the model given the parameter set $\boldsymbol{\lambda}$. The most probable model can then be determined, and the GW likelihood subsequently evaluated. Although other priors have been suggested[34,40], we used the following model prior conditional on the parameters $\boldsymbol{\lambda}$:

$$\Pi(\mathfrak{M}_i \mid \boldsymbol{\lambda}) = \frac{\mathcal{M}_i(\boldsymbol{\lambda})^{-4}}{\sum_j \mathcal{M}_j(\boldsymbol{\lambda})^{-4}}, \tag{6}$$

where $\mathcal{M}(\boldsymbol{\lambda})$ is the mismatch between the model $\mathfrak{M}_i$ and a numerical relativity simulation with parameters $\boldsymbol{\lambda}$. Equation (6) implies that the most accurate GW model will more probably be used to evaluate the likelihood in each region of the parameter space.

Although we tested several mismatch-dependent priors, equation (6) was chosen for this work because it accentuates small differences in the mismatch between models and because it was found to perform optimally. However, because the mismatch is a function of the power spectral density, it will subtly change when the profile of the power spectral density is varied, for example, due to an improvement in the sensitivity of GW detectors as a result of commissioning periods or because of small daily variations due to noise artefacts. As a result, it is possible that the relative model probabilities in equation (6) will vary for different power spectral density realizations. Note that this is a common problem in GW astronomy with, for example, search pipelines similarly using a single representative power spectral density when constructing template banks[69]. We leave a more detailed analysis investigating the choice and stability of this distribution to future work.

Constructing a match interpolant

Mismatch computations are fast, taking O(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 an appreciably longer time to evaluate, and producing the mismatch $\mathcal{M}(\boldsymbol{\lambda})$ will be a limiting cost in a Bayesian analysis because the likelihood is evaluated $O(10^8)$ times during a typical nested sampling analysis. For this reason, we constructed an interpolant for the mismatch across the parameter space, $\mathcal{M}(\boldsymbol{\lambda})$, based on a discrete set of K mismatches for each of the GW models used in this analysis.

Owing to computational limitations, we did not use numerical relativity simulations for all possible regions of the compact binary parameter space. For the aligned-spin interpolant construction, therefore, we evaluated mismatches using the numerical relativity hybrid surrogate model NRHybSur3dq8 (ref. 28) as a proxy for the numerical relativity simulations. There is a long and productive history of GW signal modelling using a variety of approaches[3,22,23,25,59,60,70–85], and we compared against the IMRPhenomXHM[83] and IMRPhenomTHM[85] waveform models, two of the leading frequency and time-domain models available for aligned-spin binaries, respectively. We did not use the state-of-the-art effective-one-body models[2,25] for the aligned-spin proof-of-principle test because IMRPhenom models are one to two orders of magnitude faster to evaluate. For the precessing model interpolants, we used the models described in 'GW Bayesian inference': IMRPhenomXPHM (ref. 22; with the updated precession formalization[26]), IMRPhenomTPHM[23] and SEOBNRv5PHM[2], and we compared the precessing models against the numerical relativity waveform surrogate model NRSur7dq4 (refs. 27,28) as a proxy for full numerical relativity simulations when computing mismatches.

We next describe, in the 'Interpolant for aligned-spin waveform mismatches' section, how we constructed an interpolant for binaries with spins aligned with the orbital angular momentum. We tested this interpolant by comparing the posterior samples obtained from a Bayesian inference analysis guided by an actual mismatch computation at every step versus a Bayesian inference analysis guided by the interpolant. We describe how we generalized this to build a generic-spin interpolant. Because of the computational cost, we used the Bayesian inference verification analysis to justify using an interpolant-guided analysis for systems with generic spins.

Interpolant for aligned-spin waveform mismatches

We begin with a test of the method using aligned-spin GW models containing higher signal multipoles. To simplify the construction of the mismatch interpolant for this test application, we reduced the dimensionality of the mismatch parameterization by artificially fixing several signal and model parameters. We chose to fix the total mass of the binary to $M = 90\,M_\odot$ and the inclination angle to $\theta_{\rm JN} = \pi/3$, where $\theta_{\rm JN}$ spans the angle between the line of sight to the binary 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 three-dimensional mismatch interpolants were constructed from mismatches computed on a uniform grid of eight points in $0.125 \leq q \leq 1$ and 17 points in each $-0.8 \leq \chi_{1,2} \leq 0.8$, providing 2,312 total mismatch points for each model. The interpolants were produced as polynomial fits to $\log_{10}\mathcal{M}$ of the form

$$\log_{10}\mathcal{M}(q,\chi_1,\chi_2) = \sum_{\substack{0\leq a\leq 6\\ 0\leq b,c\leq 8}} f_{abc}\, q^a \chi_1^{\,b}\, \chi_2^{\,c}, \tag{7}$$

with the fitting coefficients f_{abc} computed using the Fit function in Mathematica and exported to Python using FortranForm. These mismatch surfaces were well behaved, and we found that the simple-polynomial fits described provide sufficiently small relative errors (arising from equations (22) and (23) described below) of 10^{-4} and 10^{-3}, respectively, which suffices for this initial proof-of-principle test.

Next, we validated that our interpolant gives indistinguishable results compared to computing the mismatch directly in a Bayesian inference analysis. We performed two Bayesian inference analyses, both with the DYNESTY nested sampling software[47] using Bilby[48]. We used 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 used equation (7) when computing the conditional probabilities of equation (6) and in the other we directly computed the mismatch between the models and the surrogate at the sample point.

We used the Jensen–Shannon divergence to compare posterior distributions[86] because it is commonly used in GW astronomy[87,88]. The Jensen–Shannon divergence ranges between 0 bits (for statistically identical distributions) and 1 bit (for statistically distinct distributions). A general rule of thumb is that a Jensen–Shannon divergence <50 mbits implies that the distributions are in good agreement[87].

Supplementary Table 1 presents the Jensen–Shannon divergences between marginalized posterior distributions obtained when calculating the mismatch exactly and when using the interpolant. All divergences were considerably less than 50 mbits, implying that the distributions are close to statistically identical. The Bayesian analysis that used the interpolant completed in ~500 CPU hours, about ×250 faster than the Bayesian analysis that computed the mismatch exactly. Given the almost statistically identical posteriors and reduced computational cost, we used the interpolated mismatch for all subsequent analyses.

Interpolant for precessing waveform mismatches

When computing interpolants for the mismatches in equation (3), we chose to fit for the $\log_{10}$ of the sky-averaged, optimized waveform mismatch equation (3). Accordingly, our error bars become $\sigma_{\log} := |\log_{10}(\mathcal{M}-\sigma) - \log_{10}(\mathcal{M}+\sigma)|$.

Next, we generated a mismatch dataset for the fitting construction (training). We could have simply selected values for the intrinsic parameters $\{m_1, m_2, \mathbf{S}_1, \mathbf{S}_2\}$ and obtained $\mathcal{M}$ with the procedure above, but we found that the brute force use of analytic functions of eight variables to fit to this dataset was not the best approach. Instead, we opted to first reduce the dimensionality of the parameter space and then employ functional fitting. As described in Appendix A of Mac Uilliam et al.[5], we had already 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 $\boldsymbol{\lambda}_{\rm m} = \{t_c, \phi_c, \kappa, \varphi_{\rm spin}\}$, which needs to be repeated for every element of the 294-term sum in equation (3). For example, depending on the mass ratio and total mass, the computation of the average mismatch equation (3) 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.

Therefore, we had to keep in mind the computational economics when generating the data to construct the fits.

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

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

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 opted to work with the total mass, as its impact on the mismatch, equations (2) and (3), has been well documented[2,5,23,59,82].

The Cartesian components of each spin vector can 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[89]. Thus, we can write $\mathbf{S}_i = |S_i|(\sin\theta_i \cos\varphi_i, \sin\theta_i \sin\varphi_i, \cos\theta_i)^{\mathrm{T}}$ for $i = 1, 2$.

We reduced 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, which represent 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: $\{\chi_{\mathrm{p}}, \chi_{\mathrm{eff}}\}$. The former is given by[79,90]

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

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

χ_{eff} is the parallel projection counterpart to χ_{p} (refs. 79,91–93):

$$\chi_{\mathrm{eff}} = \frac{1}{1+q}(\chi_1 + q\chi_2) = \frac{1}{1+q}(\bar{S}_1 \cos\theta_1 + q\bar{S}_2 \cos\theta_2). \tag{10}$$

This is a conserved quantity up to 1.5 post-Newtonian order[92], 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 $-1 \leq \chi_{\mathrm{eff}} \leq 1$, given the Kerr spin limit $|\chi_{1,2}| \leq 1$.

Other perpendicular projections are described in the literature[61,94], but the one that we empirically determined to be the best for fitting is $\chi_\perp$ (ref. 95):

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

where $\mathbf{S}_{i,\perp} = \bar{S}_i m_i^2(\sin\theta_i \cos\phi_i, \sin\theta_i \sin\phi_i, 0)^{\mathrm{T}}$ for $i = 1, 2$. We also experimented with a generalized version of χ_{p} (ref. 96) but found this quantity to be not as well suited for fitting as χ_{p} or $\chi_\perp$. Given that the mismatches will be maximized over the in-plane spin angle φ_{spin}, we mapped $\phi_{1,2}$ 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 χ_{eff}, we introduced

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

We, thus, have several choices for each perpendicular or parallel scalar: $x = \chi_{\mathrm{p}}$ or $\chi_\perp$ and $y = \chi_{\mathrm{eff}}$ 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, however, compelled us to drop χ_{p} as it produced less faithful results, partly because it does not carry any information about the planar spin angle separation $\Delta\phi$. Thus, we were left with two possible pairings for the reduced spin space: $\{\chi_\perp, \chi_{\mathrm{eff}}\}$ and $\{\chi_\perp, \chi_\parallel\}$. Accordingly, we introduced the fitting training-set labels $K_1 = \{\chi_\perp, \chi_{\mathrm{eff}}, \eta, M\}$ 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 fitting 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 found that $\{x, y, \eta\}$ were either not correlated or weakly correlated, for which we present correlation coefficients at the end of this section.

Next, we introduced a discrete parameter grid over the chosen four-dimensional $\{x, y, \eta, M\}$ space, which we used for fitting. We limited η to the range from 0.16 (corresponding to $q = 1/4$) to 0.25 ($q = 1$) in four even steps, resulting in five distinct values η_j, $j = 1, \ldots, 5$. For the total mass, we employed $M = \{75, 117.5, 150\}\,M_\odot$ as our grid points, chosen because of the following reasons. (1) NRSur7dq4 was trained with data only from binaries with $q \geq 1/4$. (2) The time length limit in NRSur7dq4 (ref. 27) of $4{,}300M$ imposes $M \gtrsim 75\,M_\odot$ in order for the binary to enter the detector bandwidth at a GW frequency of 20 Hz. (3) Binaries with $M > 150\,M_\odot$ mostly emit merger-ringdown signals in the detection band[5], thus leaving hardly any imprint of precession in the waveform reconstructed from detector data. (4) Model mismatches tend to weakly depend on the total mass[2,5,23,59,82], thus three grid points in mass space suffice for our current purposes given the computational burden of generating new data.

For better fitting performance, the remaining two fitting parameters, x and y, also had to be placed on a regular grid. However, the quantities that we picked to cover this space, namely the pairings $\{\chi_\perp, \chi_{\mathrm{eff}}\}$ and $\{\chi_\perp, \chi_\parallel\}$, are not intrinsic parameters of the binary system. To construct a regular grid in $\{x, y\}$, therefore, we started from a regular grid of roughly 50,000 elements in $\{\bar{S}_1, \bar{S}_2, \theta_1, \theta_2, \Delta\phi\}$ space and used this to populate the $\{x, y\}$ space with values of q already determined by the η_j grid. The resulting grid in, for example, the $\chi_\perp$–χ_{eff} plane, is show as a scatter plot of blue dots in the left panel of Supplementary Fig. 4. The parameter space seems to be bounded by a half prolate ellipse drawn as the orange curve. The horizontal and vertical axes of the ellipse are given by

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

Guided by this observation, we constructed a regular, elliptical grid in $\{x, y\}$ space as follows. First, we introduced 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 $A\sinh\mu = a$ and $A\cosh\mu = b$. Note that in equations (14a)–(15), we swapped $\cosh\mu$ and $\sinh\mu$ because, as we show below, our ellipses are prolate, that is, $a < b$.

Here, we aimed to create a grid based on 'concentric' ellipses with the same aspect ratio, starting with the outermost one (orange curve in the left panel of Supplementary Fig. 4). With A and μ fixed, we created an elliptical grid of our choosing:

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

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

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 Supplementary Fig. 4 represented by the red dots with $N_r = 10$ and $N_s = 24$, that is, 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 that we sought had to be chosen such that the corresponding values for $\{x, y\}$ yield points on the elliptical grid, that is, the red dots in the left panel of Supplementary Fig. 4. We started 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 kth grid point with coordinates $\{x_k, y_k\}$, we found 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 used these values as initial guesses in a root-finding 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 q values used are consistent with our aforementioned η_j grid.

As this is numerical root-finding, we replaced the right-hand sides of equations (17a) and (17b) with a threshold of 10^{-12}. We ran this root-finding procedure for every single elliptical grid point. The result is shown in the middle panel of Supplementary Fig. 4. Over each red dot, we have placed a faint blue dot representing the grid points that our algorithm found. On average, each numerically determined grid point was offset by $\leq 10^{-12}$ from the exact grid (red) point. For a grid of 250 points, this amounted to a total grid offset of $\lesssim 3 \times 10^{-9}$. We actually found this number to be 1.5×10^{-8} for the elliptical $\{\chi_\perp, \chi_{\rm eff}\}$ grid of Supplementary Fig. 4 because we had to relax our strict tolerance from 10^{-12} to 10^{-10} for certain grid points to speed up the procedure. As we show further below, a grid offset of $\sim 10^{-8}$ is much smaller than the average fit unfaithfulness that we obtained, $\sim \mathcal{O}(10^{-2})$, and, thus, was 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}. Note that a few of the intrinsic coordinates for the grid points exceed the training limit for NRSur7dq4 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. Therefore, we went one step further and transformed the elliptical coordinates into rectangular ones:

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

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

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

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

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

Comparing equations (16a) with (18a) and (16b) with (18b) gives the $N_r \times N_s$ rectangular grid $\{X_r, Y_s\}$ with $r = 1, \ldots, N_r$ and $s = 0, \ldots, N_s$, which we show in the right panel of Supplementary Fig. 4. 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 introduced the rescaled variables $Z = 4\eta$ and $V = M/(75\,M_\odot)$.

After much trial and error, we settled on the following fitting function:

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

We chose this particular form to better curb the extrapolation behaviour of the fitting in parts of the $\{Z, V\}$ (mass ratio, total mass) space outside the training region $Z < 0.64$ ($\eta < 0.16$) and $V < 1 \cup V > 2$ corresponding to $M < 75\,M_\odot \cup M > 150\,M_\odot$. We used two-dimensional polynomials in the $\{X, Y\}$ subspace of the fitting 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 values of $\{n_i, n_j, n_k\}$ in the triple summation of equation (20) were chosen such that we had at most roughly the same number of fitting parameters as the total number of grid points used in the $\{x, y, \eta\}$ subspace, which for Supplementary Fig. 4, for example, was $10 \times 25 = 250$. Note that in the denominator of equation (20), we used the absolute value of the fitting coefficients c_{ijkl} to ensure that there were no singularities. We also set $c_{ij02} = 1$, which is the leading term in the denominator, a standard choice 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 $N = N_r(N_s + 1)$ 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, for example, 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 the NonlinearModelFit function in Mathematica and store the coefficients $c_{ijkl,L}$ of equation (20).

We started our fitting optimization routine with $\{n_i, n_j, n_k\} = \{4, 3, 2\}$ and generated fits up to some $\{n_i^{\max}, n_j^{\max}, n_k^{\max}\}$ to ensure that the total number of fitting parameters was ~N. The choice of $\{4, 3, 2\}$ yielded 160 fitting parameters. Smaller values of $\{n_i, n_j, n_k\}$ resulted in fewer than 100 fitting coefficients and led to underfitting for training grids of size $\gtrsim \mathcal{O}(200)$, which, as we explain below, is the grid size that we adopted. Our routine picked as the final fit the one for which the values of $\{n_i, n_j, n_k\}$ in equation (20) yielded the lowest relative difference with respect to the training dataset. For this purpose, we defined the relative difference between the data and the fit at the kth point:

$$\Delta_{\rm rel}^k := 1 - \frac{\mathcal{F}(X_k, Y_k, Z_k, V_k)}{\log_{10}\mathcal{M}_k}, \tag{21}$$

and introduced two quantities to gauge fit quality during training. The first is the l^2-norm of $\Delta_{\rm rel}^k$ between the fit and the data normalized by the length of the vector:

$$\Delta_{\rm rel}^{(1)} := \frac{1}{3N} \sqrt{\sum_{k=1}^{3N} \left|\Delta_{\rm rel}^k\right|^2}, \tag{22}$$

and the second is the signed average relative difference:

$$\Delta_{\rm rel}^{(2)} := \frac{1}{3N} \sum_{k=1}^{3N} \left(\Delta_{\rm rel}^k\right), \tag{23}$$

which tells us whether the fit globally over- or underestimates the data.

We picked values for $\{n_i, n_j, n_k\}$ that simultaneously minimized both of the above relative differences. These relative differences are the most important fitting attributes for this work, as we must robustly predict the mismatches to numerical relativity when selecting the appropriate model to use at a given point in the parameter space. When more than one set of values for $\{n_i, n_j, n_k\}$ was returned, we opted for the set that yielded a reduced chi squared (χ^2/d.f.) closest to unity. Once the fitting training was complete with the above optimization of $\{n_i, n_j, n_k\}$, we checked the fitting performance over an appropriate verification set, which we discuss further below.

The question of training grid resolution can be answered only after setting an unfaithfulness threshold for the target fit. We aimed for $\Delta^{|\mathrm{av}|}_{\mathrm{rel}} \approx 0.05$ for each fit, where

$$\Delta^{|\mathrm{av}|}_{\mathrm{rel}} := \frac{1}{3N}\sum_{k=1}^{3N} |\Delta^{k}_{\mathrm{rel}}| \tag{24}$$

is the average absolute relative disagreement between the fit and the verification data. With the above threshold established, we set out to determine whether the $\{x, y\}$ training grid of size 10×25 sufficed. First, we downsampled this grid to create coarser grids of dimension 10×13, 5×13 and 5×7 and computed $\Delta^{|\mathrm{av}|}_{\mathrm{rel}}$ for each with respect to the original verification set (of size 250). As we gradually increased the grid size from 5×7 to 10×25, we observed $\Delta^{|\mathrm{av}|}_{\mathrm{rel}}$ decreasing from ~0.10 to $\lesssim$0.05 for the fits listed in Supplementary Table 2. For example, the fit used to make Supplementary Fig. 5 yielded $\Delta^{|\mathrm{av}|}_{\mathrm{rel}} = 0.048$. Increasing the grid size to $\mathcal{O}(1{,}000)$ elements should further reduce $\Delta^{|\mathrm{av}|}_{\mathrm{rel}}$. However, this quickly turns into a problem of diminishing returns given that it would take one month to generate the mismatch data using 128 CPUs on this grid.

Furthermore, an inspection of the structure of the mismatch data revealed that an elliptical grid with $\mathcal{O}(10)$ points along the radial direction and $\mathcal{O}(20)$ points along the azimuthal direction sufficed to capture the dominant trends in the data at the level of fitting unfaithfulness that we sought: $\Delta^{|\mathrm{av}|}_{\mathrm{rel}} \approx 0.05$. Thus, our aforementioned grid of dimensions 10×25 had sufficient resolution. We leave fitting improvements to future work, which we have already begun undertaking.

Supplementary Fig. 5 shows contour plots of the unfaithfulness of the fit for the $\log_{10}$ of the NRSur7dq4 versus SEOBNRv5PHM mismatches to the verification dataset. Because of the computational burden required to obtain the mismatches, we used data with $y = \chi_{\parallel}$ (χ_{eff}) to verify the data trained with χ_{eff} ($\chi_{\parallel}$). This resulted in verification sets that were the same size as the training sets, so ours is rather a harsh verification test. The contours represent the absolute value of the relative difference between the fit and the data. The fitting was 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 was 0.05 or less. Note that this quantity is not $\Delta^{(2)}_{\mathrm{rel}}$ applied to the verification set but rather the absolute value of the summand in equation (23). The fits for the NRSur7dq4 versus IMRPhenomTPHM and the NRSur7dq4 versus IMRPhenomXPHM mismatches also yielded similar level of agreement, as did the fits trained with the $y = \chi_{\mathrm{eff}}$ set.

We summarize these results and provide other metrics for all the fits in Supplementary Table 2. The average relative distance (equation (22)) between each fit and the corresponding verification data was always less than 4×10^{-3}, and the average relative difference (equation (23)) had magnitude less than 0.02. Interestingly, $\Delta^{(2),\mathrm{ver}}_{\mathrm{rel}}$ for the verification set was negative for most fits, indicating that the fits slightly overestimate the data. The value of $1 - \bar{R}^2$ was $\lesssim$0.01 for all our fits, where $\bar{R}^2$ is the reduced R squared. For most cases, we observed χ^2/d.f. $\approx \mathcal{O}(1)$. The cases for which this quantity was about an order of magnitude lower arose because we overestimated 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 took to be our individual data points.

As another check of the fits, we investigated their behaviour in the extrapolation region corresponding to $\eta < 0.16$ ($q < 1/4$), $\bar{S}_{1,2} > 0.8$ and $M < 75\,M_{\odot} \cup M > 150\,M_{\odot}$. Recall that we chose the particular form of equation (20) for the fitting function to better control unwanted extrapolation behaviour, such as the 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 very distant from the training (interpolation) regime. On the other hand, because the relevant two-dimensional cut of the training region covered most of the $\{x, y\}$ space, polynomial extrapolation should not cause issues.

We illustrate all this in Supplementary Fig. 6, where we plot the fit in equation (20) to the $\log_{10}$ of the NRSur7dq4 versus 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 $\{\eta, \bar{S}_{1,2}\} = \{0.139, 0.85\}$ ($q = 1/5$), and other intrinsic parameters were chosen accordingly. Similarly, the red ellipse traces the $\{\eta, \bar{S}_{1,2}\} = \{0.122, 0.9\}$ ($q = 1/6$) set and the orange ellipse the edge of the fitting training region with $\{\eta, \bar{S}_{1,2}\} = \{0.16, 0.8\}$ ($q = 1/4$), which was shown in Supplementary Fig. 4. 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 $M = 50$ to $200\,M_{\odot}$, but at each separate elliptical coordinate, hence the blue, red and orange curves. The shaded grey region in each inset marks the training range $M \in [75, 150]\,M_{\odot}$ for the fits, which is actually relevant only to the orange curve, as the blue and the red curves are, by definition, outside the training region. Thanks to the specific functional form of the fit, the extrapolation does not exhibit any pathologies. Additionally, note that the blue curves mostly lie 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 $q \geq 1/5$ and $\bar{S}_i \leq 0.85$.

As a further test of the fitting performance in its extrapolation regime, we performed two more parameter recoveries of a numerical relativity simulation, SXS:BBH:1156, with the injected value for M set to 75 or $100\,M_{\odot}$ and $q = 0.228$, placing the former 'squarely' outside the $M > 75\,M_{\odot}$ and $q \geq 1/4$ training regime of our fits. This simulation also had $|\mathbf{S}_{2,\perp}|/m_2^2 \approx 0.76$ On the other hand, because whereas the primary had negligible planar spin. We show the results of our method applied to this Bayesian inference analysis in Supplementary Fig. 3. The two-dimensional posteriors from the $M = 75\,M_{\odot}$ ($100\,M_{\odot}$) injections are plotted in the top (bottom) panels. In both analyses, our method recovered the injected values for m_1 and m_2 and the effective spins very well, with most samples clustered near the injected values. Such a good recovery of the masses and, hence, the mass ratio for the $M = 75\,M_{\odot}$ injection is an indirect testament to the robustness of the fit (equation (20)), especially because slightly more than half of the total mass posteriors for this particular Bayesian analysis happened to be less than $75\,M_{\odot}$.

We conclude this section by briefly returning to two issues. The first is that the fitting variables $\{\eta, x, y\}$ used to construct the fit are not fully independent of each other, as each is a function of the mass ratio q. The second regards the choice of power spectral density used when calculating the mismatch and, hence, the fits. For the first issue, as we have explained already, the q-scaled spin projections yielded 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, $K_1(y = \chi_{\mathrm{eff}})$ and $K_2(y = \chi_{\parallel})$. For K_1, we obtained $\mathcal{C}_{\chi_{\perp}-\chi_{\mathrm{eff}}} = 0.09$, $\mathcal{C}_{\eta-\chi_{\mathrm{eff}}} = 0.02$ and $\mathcal{C}_{\eta-\chi_{\perp}} = -0.315$. For K_2, we have $\mathcal{C}_{\chi_{\perp}-\chi_{\parallel}} = -0.1$, $\mathcal{C}_{\eta-\chi_{\perp}} = 0.03$ and $\mathcal{C}_{\eta-\chi_{\parallel}} = 0.216$. For both parameter sets used in fitting training, we have either uncorrelated pairings of fitting variables or weakly correlated pairings. That $|\mathcal{C}_{\eta-\chi_{\parallel}}|$ of set K_2 is less than $|\mathcal{C}_{\eta-\chi_{\perp}}|$ of set K_1 may partly explain

why we observed slightly better performance from the fits constructed from K_2, as indicated in Supplementary Table 2.

For the second issue, as already explained, we used the design sensitivity of Advanced LIGO[46] when computing the mismatch, which subtly changes when the profile of the power spectral density is varied. If this method were to be used during live observing run periods, during which the power spectral density changes every day owing to noise artefacts in the GW strain data, we would suggest using a harmonic average power spectral density estimated from engineering run data to calculate mismatches, as is commonly done in GW search pipelines[69]. The fit would then be reconstructed before each GW observing run.

Data availability

The aligned-spin and generic-spin match interpolants as well as the posterior samples from the analyses performed in this work are available at https://icg-gravwaves.github.io/incorporating_model_uncertainty_into_Bayesian_inference.

Code availability

Python scripts detailing our modifications to Bilby[48] are available at https://icg-gravwaves.github.io/incorporating_model_uncertainty_into_Bayesian_inference.

References

1. Veitch, J. et al. Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library. *Phys. Rev. D* **91**, 042003 (2015).
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).
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).
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. *Phys. Rev. D* **110**, 064024 (2024).
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).
6. Dhani, A. et al. Systematic biases in estimating the properties of black holes due to inaccurate gravitational-wave models. Preprint at https://doi.org/10.48550/arXiv.2404.05811 (2024).
7. Akcay, S. et al. Waging a campaign: results from an injection-recovery study involving 35 numerical relativity simulations and three waveform models. Preprint at https://doi.org/10.48550/arXiv.2506.19990 (2025).
8. Pürrer, M. & Haster, C.-J. Gravitational waveform accuracy requirements for future ground-based detectors. *Phys. Rev. Res.* **2**, 023151 (2020).
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).
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).
11. Aasi, J. et al. Advanced LIGO. *Class. Quantum Gravity* **32**, 074001 (2015).
12. Acernese, F. et al. Advanced Virgo: a second-generation interferometric gravitational wave detector. *Class. Quantum Gravity* **32**, 024001 (2014).
13. Akutsu, T. et al. Overview of KAGRA: detector design and construction history. *Prog. Theor. Exp. Phys.* **2021**, 05A101 (2021).
14. Hamilton, E. et al. Catalog of precessing black-hole-binary numerical-relativity simulations. *Phys. Rev. D* **109**, 044032 (2024).
15. Mroue, A. H. et al. A catalog of 174 binary black-hole simulations for gravitational-wave astronomy. *Phys. Rev. Lett.* **111**, 241104 (2013).
16. Boyle, M. et al. The SXS Collaboration catalog of binary black hole simulations. *Class. Quantum Gravity* **36**, 195006 (2019).
17. Healy, J., Lousto, C. O., Zlochower, Y. & Campanelli, M. The RIT binary black hole simulations catalog. *Class. Quantum Gravity* **34**, 224001 (2017).
18. 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).
19. Healy, J. & Lousto, C. O. Third RIT binary black hole simulations catalog. *Phys. Rev. D* **102**, 104018 (2020).
20. Healy, J. & Lousto, C. O. Fourth RIT binary black hole simulations catalog: extension to eccentric orbits. *Phys. Rev. D* **105**, 124010 (2022).
21. Jani, K. et al. Georgia Tech catalog of gravitational waveforms. *Class. Quantum Gravity* **33**, 204001 (2016).
22. 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).
23. Estellés, H. et al. New twists in compact binary waveform modeling: a fast time-domain model for precession. *Phys. Rev. D* **105**, 084040 (2022).
24. Albertini, A., Nagar, A., Rettegno, P., Albanesi, S. & Gamba, R. Waveforms and fluxes: towards a self-consistent effective one body waveform model for nonprecessing, coalescing black-hole binaries for third generation detectors. *Phys. Rev. D* **105**, 084025 (2022).
25. 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).
26. Colleoni, M., Vidal, F. A. R., García-Quirós, C., Akçay, S. & Bera, S. Fast frequency-domain gravitational waveforms for precessing binaries with a new twist. *Phys. Rev. D* **111**, 104019 (2025).
27. Varma, V. et al. Surrogate models for precessing binary black hole simulations with unequal masses. *Phys. Rev. Res.* **1**, 033015 (2019).
28. Varma, V. et al. Surrogate model of hybridized numerical relativity binary black hole waveforms. *Phys. Rev. D* **99**, 064045 (2019).
29. Owen, B. J. Search templates for gravitational waves from inspiraling binaries: choice of template spacing. *Phys. Rev. D* **53**, 6749–6761 (1996).
30. Abbott, R. et al. GWTC-3: compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run. *Phys. Rev. X* **13**, 041039 (2023).
31. Ashton, G. & Khan, S. Multiwaveform inference of gravitational waves. *Phys. Rev. D* **101**, 064037 (2020).
32. 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).
33. Ashton, G. & Dietrich, T. The use of hypermodels to understand binary neutron star collisions. *Nat. Astron.* **6**, 961–967 (2022).
34. Hoy, C. Accelerating multimodel Bayesian inference, model selection, and systematic studies for gravitational wave astronomy. *Phys. Rev. D* **106**, 083003 (2022).
35. Moore, C. J. & Gair, J. R. Novel method for incorporating model uncertainties into gravitational wave parameter estimates. *Phys. Rev. Lett.* **113**, 251101 (2014).
36. Doctor, Z., Farr, B., Holz, D. E. & Pürrer, M. Statistical gravitational waveform models: what to simulate next? *Phys. Rev. D* **96**, 123011 (2017).
37. Williams, D., Heng, I. S., Gair, J., Clark, J. A. & Khamesra, B. Precessing numerical relativity waveform surrogate model for binary black holes: a Gaussian process regression approach. *Phys. Rev. D* **101**, 063011 (2020).

38. Read, J. S. Waveform uncertainty quantification and interpretation for gravitational-wave astronomy. *Class. Quantum Gravity* **40**, 135002 (2023).
39. Khan, S. Probabilistic model for the gravitational wave signal from merging black holes. *Phys. Rev. D* **109**, 104045 (2024).
40. Jan, A. *Marginaliz*https://doi.org/10.48550/arXiv.2506.19990*ing over the Waveform Systematics of Compact Binary Coalescence Models using RIFT*. Master's thesis, Rochester Institute of Technology (2021).
41. Blackman, J. et al. Numerical relativity waveform surrogate model for generically precessing binary black hole mergers. *Phys. Rev. D* **96**, 024058 (2017).
42. 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. D* **49**, 6274–6297 (1994).
43. Fairhurst, S., Green, R., Hoy, C., Hannam, M. & Muir, A. Two-harmonic approximation for gravitational waveforms from precessing binaries. *Phys. Rev. D* **102**, 024055 (2020).
44. Hoy, C., Fairhurst, S. & Mandel, I. Rarity of precession and higher-order multipoles in gravitational waves from merging binary black holes. *Phys. Rev. D* **111**, 023037 (2025).
45. Goldberg, J. N., MacFarlane, A. J., Newman, E. T., Rohrlich, F. & Sudarshan, E. C. G. Spin-s spherical harmonics and ð. *J. Math. Phys.* **8**, 2155–2166 (1967).
46. LIGO Scientific Collaboration and Virgo Collaboration. Noise curves used for simulations in the update of the observing scenarios paper. *DCC* https://dcc.ligo.org/LIGO-T2000012/public (2022).
47. Speagle, J. S. DYNESTY: a dynamic nested sampling package for estimating Bayesian posteriors and evidences. *Mon. Not. R. Astron. Soc.* https://doi.org/10.1093/mnras/staa278 (2020).
48. Ashton, G. et al. Bilby: a user-friendly Bayesian inference library for gravitational-wave astronomy. *Astrophys. J. Suppl. Ser.* **241**, 27 (2019).
49. Pürrer, M., Hannam, M. & Ohme, F. Can we measure individual black-hole spins from gravitational-wave observations? *Phys. Rev. D* **93**, 084042 (2016).
50. Ajith, P. et al. Inspiral-merger-ringdown waveforms for black-hole binaries with non-precessing spins. *Phys. Rev. Lett.* **106**, 241101 (2011).
51. Ajith, P. Addressing the spin question in gravitational-wave searches: waveform templates for inspiralling compact binaries with nonprecessing spins. *Phys. Rev. D* **84**, 084037 (2011).
52. Baird, E., Fairhurst, S., Hannam, M. & Murphy, P. Degeneracy between mass and spin in black-hole-binary waveforms. *Phys. Rev. D* **87**, 024035 (2013).
53. Toubiana, A. & Gair, J. R. Indistinguishability criterion and estimating the presence of biases. Preprint at https://doi.org/10.48550/arXiv.2401.06845 (2024).
54. Hannam, M. et al. General-relativistic precession in a black-hole binary. *Nature* **610**, 652–655 (2022).
55. Woosley, S. E., Heger, A. & Weaver, T. A. The evolution and explosion of massive stars. *Rev. Mod. Phys.* **74**, 1015–1071 (2002).
56. Islam, T. et al. Analysis of GWTC-3 with fully precessing numerical relativity surrogate models. Preprint at https://doi.org/10.48550/arXiv.2309.14473 (2023).
57. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
58. Harry, I., Privitera, S., Bohé, A. & Buonanno, A. Searching for gravitational waves from compact binaries with precessing spins. *Phys. Rev. D* **94**, 024012 (2016).
59. 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).
60. 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. D* **100**, 024059 (2019).
61. Hamilton, E. et al. Model of gravitational waves from precessing black-hole binaries through merger and ringdown. *Phys. Rev. D* **104**, 124027 (2021).
62. Bruegmann, B., Gonzalez, J. A., Hannam, M., Husa, S. & Sperhake, U. Exploring black hole superkicks. *Phys. Rev. D* **77**, 124047 (2008).
63. Keppel, D., Nichols, D. A., Chen, Y. & Thorne, K. S. Momentum flow in black hole binaries. I. Post-Newtonian analysis of the inspiral and spin-induced bobbing. *Phys. Rev. D* **80**, 124015 (2009).
64. Kalaghatgi, C. & Hannam, M. Investigating the effect of in-plane spin directions for precessing binary black hole systems. *Phys. Rev. D* **103**, 024024 (2021).
65. Ramos-Buades, A., Schmidt, P., Pratten, G. & Husa, S. Validity of common modeling approximations for precessing binary black holes with higher-order modes. *Phys. Rev. D* **101**, 103014 (2020).
66. Ghosh, S., Kolitsidou, P. & Hannam, M. First frequency-domain phenomenological model of the multipole asymmetry in gravitational-wave signals from binary-black-hole coalescence. *Phys. Rev. D* **109**, 024061 (2024).
67. Kolitsidou, P., Thompson, J. E. & Hannam, M. Impact of antisymmetric contributions to signal multipoles in the measurement of black-hole spins. *Phys. Rev. D* **111**, 024050 (2025).
68. Skilling, J. Nested sampling for general Bayesian computation. *Bayesian Anal.* **1**, 833–859 (2006).
69. Dal Canton, T. & Harry, I. W. Designing a template bank to observe compact binary coalescences in Advanced LIGO's second observing run. Preprint at https://doi.org/10.48550/arXiv.1705.01845 (2017).
70. 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. D* **95**, 044028 (2017).
71. Cotesta, R. et al. Enriching the symphony of gravitational waves from binary black holes by tuning higher harmonics. *Phys. Rev. D* **98**, 084028 (2018).
72. 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).
73. Ossokine, S. et al. Multipolar effective-one-body waveforms for precessing binary black holes: construction and validation. *Phys. Rev. D* **102**, 044055 (2020).
74. Babak, S., Taracchini, A. & Buonanno, A. Validating the effective-one-body model of spinning, precessing binary black holes against numerical relativity. *Phys. Rev. D* **95**, 024010 (2017).
75. Pan, Y. et al. Inspiral-merger-ringdown waveforms of spinning, precessing black-hole binaries in the effective-one-body formalism. *Phys. Rev. D* **89**, 084006 (2014).
76. Husa, S. et al. Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal. *Phys. Rev. D* **93**, 044006 (2016).
77. Khan, S. et al. Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era. *Phys. Rev. D* **93**, 044007 (2016).
78. London, L. et al. First higher-multipole model of gravitational waves from spinning and coalescing black-hole binaries. *Phys. Rev. Lett.* **120**, 161102 (2018).
79. Hannam, M. et al. Simple model of complete precessing black-hole-binary gravitational waveforms. *Phys. Rev. Lett.* **113**, 151101 (2014).
80. 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. D* **98**, 104052 (2018).

81. Khan, S., Ohme, F., Chatziioannou, K. & Hannam, M. Including higher order multipoles in gravitational-wave models for precessing binary black holes. *Phys. Rev. D* **101**, 024056 (2020).
82. 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).
83. 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).
84. 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).
85. 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).
86. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **37**, 145–151 (1991).
87. 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. X* **9**, 031040 (2019).
88. 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).
89. Schmidt, P., Harry, I. W. & Pfeiffer, H. P. Numerical relativity injection infrastructure. Preprint at https://doi.org/10.48550/arXiv.1703.01076 (2017).
90. 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. D* **91**, 024043 (2015).
91. Damour, T. Coalescence of two spinning black holes: an effective one-body approach. *Phys. Rev. D* **64**, 124013 (2001).
92. Racine, E. Analysis of spin precession in binary black hole systems including quadrupole-monopole interaction. *Phys. Rev. D* **78**, 044021 (2008).
93. 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).
94. 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).
95. 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).
96. Gerosa, D. et al. A generalized precession parameter χ_p to interpret gravitational-wave data. *Phys. Rev. D* **103**, 064067 (2021).

Acknowledgements

We thank A. Yelikar for comments made during the LIGO–Virgo–KAGRA internal review, as well as C. Berry, A. Göttel, M. Isi, L. Thomas, M. Williams and A. Zimmerman for discussions during our presentation to the LIGO–Virgo–KAGRA Collaboration. We are also grateful to M. Hannam and L. Nuttall for discussions throughout this project. We thank A. Chua and the California Institute of Technology for hosting C.H. and S.A. in March 2024, which gave the authors time to discuss and develop this project. C.H. thanks the UKRI Future Leaders Fellowship for support (grant MR/T01881X/1). S.A. and J.M.U. acknowledge support from the University College Dublin Ad Astra Fellowship, and J.E.T. acknowledges support from a NASA LISA Preparatory Science Grant (20-LPS20-0005). This work used the computational resources provided by the ICG, SEPNet and the University of Portsmouth, which is supported by the STFC (grant ST/N000064). This research made use of data, software and/or web tools obtained from the Gravitational Wave Open Science Center (https://www.gw-openscience.org), a service of LIGO Laboratory, the LIGO Scientific Collaboration and the Virgo Collaboration. LIGO is funded by the US National Science Foundation (NSF). Virgo is funded by the French Centre National de Recherche Scientifique, the Italian Istituto Nazionale della Fisica Nucleare and the Dutch Nikhef with contributions from Polish and Hungarian institutes. This material is based upon work supported by NSF's LIGO Laboratory, which is a major facility fully funded by the NSF.

Author contributions

C.H. conceptualized the idea of sampling over several models with priors dictated by their mismatch to numerical relativity simulations as a method for incorporating model uncertainty into GW Bayesian inference. S.A. and J.E.T. developed the idea of building mismatch interpolants for this application. S.A. and J.E.T. initiated and formed the project team. C.H. implemented the method presented here into Bilby (ref. 48) and performed all the parameter estimation analyses. J.E.T. generated the aligned-spin mismatches and the aligned-spin interpolant. S.A. constructed the generic-spin interpolant. C.H. and J.E.T. investigated choices for the model conditional prior. J.M.U. generated most of the generic-spin mismatches, with C.H. and S.A. generating a subset. All authors contributed to the interpretation of the results and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41550-025-02579-7.

Correspondence and requests for materials should be addressed to Charlie Hoy.

Peer review information *Nature Astronomy* thanks Zack Carson, Yi-Ming Hu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.