Incorporation of model accuracy in 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. † e-mail: charlie.hoy@port.ac.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 further apply our method to a real gravitational-wave signal and, when assuming the binary black hole hypothesis, demonstrate that the source of GW191109_010717 has unequal component masses, with the primary having a 69% probability that it lies above the maximum black hole mass from stellar collapse. 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[2–6] (MacUilliam et al., in prep.), 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[7–9]. 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[10–12]).

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[13]. For this reason only several thousand simulations are currently available[13–20]. As a result, the latest GW models rely on analytic or semi-analytical prescriptions that are calibrated to the numerical relativity simulations[2,3,21–25], or are based on surrogate modelling techniques[26,27]. However, each modelling approach will incur some degree of approximation errors.

The accuracy of a GW model is typically measured by the mismatch[28] between the model and a fiducial waveform, often a numerical relativity simulation. The mismatch varies between 0, signify-

ing 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[29] or according to their relative Bayesian evidence[30], or by averaging the likelihood[31]. An alternative technique involves sampling over a set of GW models in a single *joint* Bayesian analysis[32,33]. 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[34–38]. These methods have either not been demonstrated in practice, or are only suitable for a single model. Preliminary work has investigated incorporating model accuracy into likelihood averaging techniques for simplified models[39]. However, this approach maintains a comparable computational cost to evidence mixing[30] and is difficult to interface with standard Bayesian inference techniques.

In this work we present the first approach to incorporate the accuracy of multiple cutting-edge models into a single GW Bayesian analysis, while also reducing the computational cost (see Methods). This technique accounts for modelling errors by prioritising 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[10]–Virgo[11]–KAGRA[12] GW detectors, we demonstrate that current techniques are more likely to 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 an example of a theoretical GW signal expected from general relativity, specifically, the `SXS:BBH:0926`[15,40] 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 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[41] is significant, and contributes a signal-to-noise ratio[42] ~ 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] (MacUilliam et al., in prep.) Such a system with significant spin-induced orbital precession has been predicted to be observed once in every 50 GW observations made by the LIGO, Virgo and KAGRA
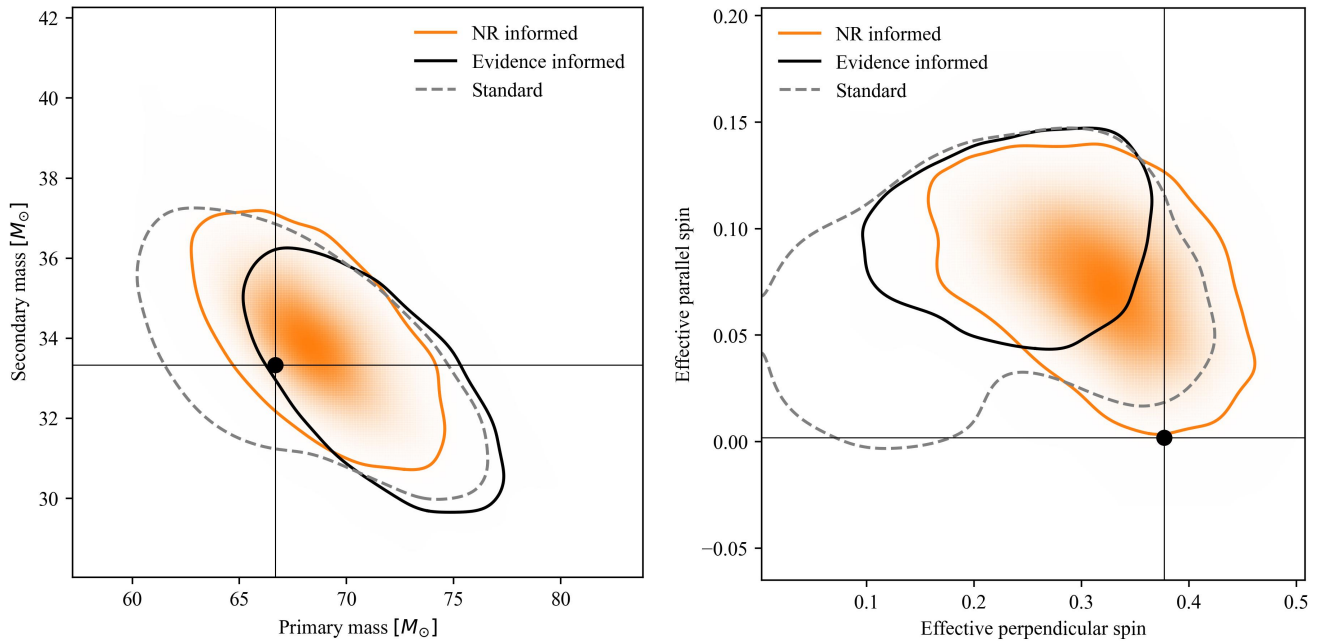
Figure 1 | Two-dimensional posterior probabilities obtained in our analysis of the `SXS:BBH:0926` numerical relativity simulation. The left panel shows the measurement of the primary and secondary mass of the binary, and the right panel shows the inferred effective parallel and perpendicular spin components (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. The contours represent 90% credible intervals and the black cross hairs indicate the true value.

GW observatories based on current black hole population estimates[43].

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[21] (with the updated precession formulation[25]), IMRPHENOMTPHM[22] and SEOBNRv5PHM[2]. All models include the general relativistic phenomenon of spin-induced orbital precession[41] and higher order multipole moments[44]. We analyse 8 seconds of data, and only consider frequencies between [20, 2048] Hz. We generate the numerical relativity simulation from ≈ 10 Hz to ensure that most higher multipole content is generated prior to our analysis window. Our analysis is restricted to a two-detector network of LIGO-Hanford and LIGO-Livingston[10], and we assume a theoretical power spectral density for Advanced LIGO's design sensitivity[45]. We use the most agnostic priors available for all parameters, identical to those used in all detections made by the LIGO–Virgo–KAGRA collaboration[29]: flat in the component masses, spin magnitudes and cosine of the spin tilt angles. We perform Bayesian inference with the `Dynesty` Nested sampling software[46] via `Bilby`[47], as has been done in all LIGO–Virgo–KAGRA analyses since the third GW catalog[29].

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 informed* combines separate inference analyses obtained with different GW models according to their relative Bayesian evidence[30], and *Standard* combines the results of separate inference analysis with equal weight. Standard is the currently adopted method by the LIGO–Virgo–KAGRA collaboration[29] as it is likely the most agnostic. 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 uncer-

tainty is inflated in comparison to the method presented here and to the Evidence informed 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[48], we consider the measurement of effective spin parameters that describe the dominant spin effects of the observed GW signal[49,50]. 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[30], 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[33]. Our analysis, on the other hand, predominantly uses SEOBNRv5PHM; specifically 90% of the time, while 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 parameter space, there is no guarantee that this model is accurate enough to avoid biases in the inferred parameter estimates[51,52]. 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 results in this region. Specfically, it yields mismatches $\sim 3\times$ and $\sim 1.8\times$ smaller than IMRPHENOMXPHM and IMRPHENOMTPHM, respectively.
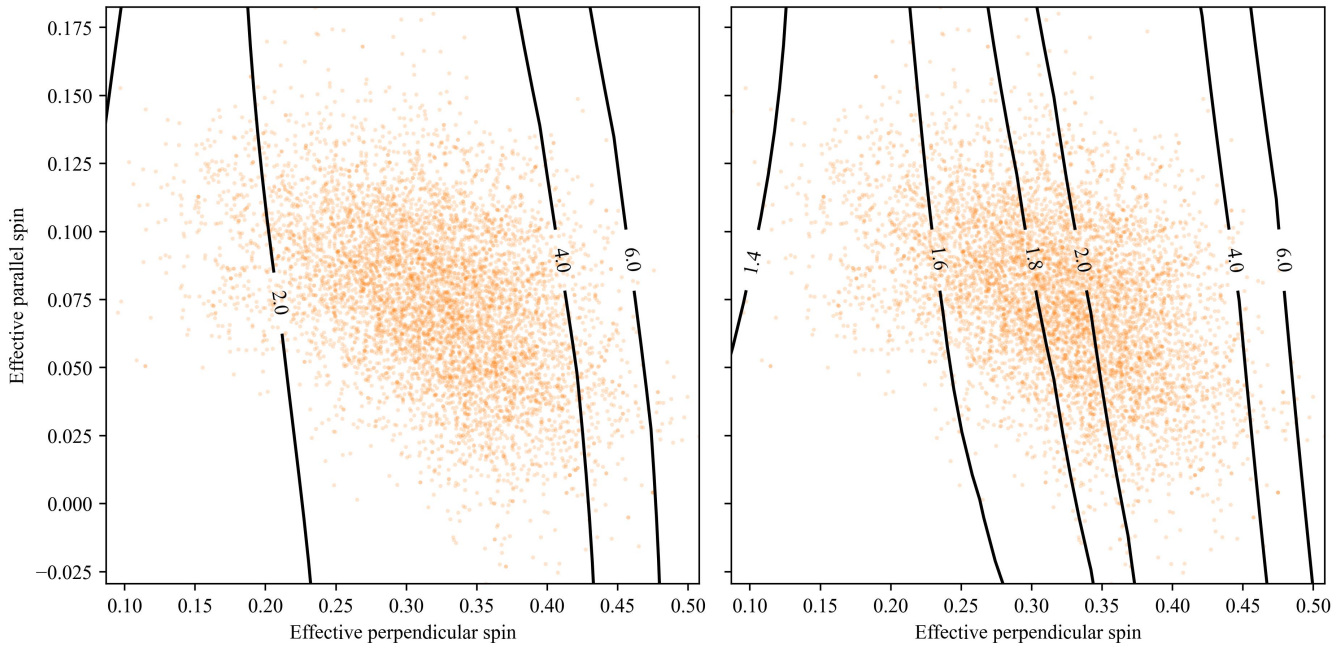
Figure 2 | Ratio of mismatches to numerical relativity simulations. Contour plot showing the ratio of mismatches to numerical relativity simulations for different effective parallel and perpendicular spin components when averaging over different mass configurations. The left panel compares IMRPHENOMXPHM 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.

Since the NR informed approach chooses the GW model based on its accuracy to numerical relativity in each region of 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 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, IMRPHENOMT-PHM 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. If SEOBNRv5PHM were removed from this analysis, we find 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 parameter space.

A single analysis with the model that is, on average, the most accurate in the parameter space of interest can be performed[53]. 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 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 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 multiple models into a single GW Bayesian analysis fully utilises this information.

Numerical relativity surrogate techniques provide accurate models for describing GWs produced from colliding black holes[26,27]. We do not sample over surrogate models in this work since they are used as a proxy for numerical relativity simulations to assess model accuracy (see Methods). We quantify the efficacy of our approach by comparing results to those obtained with surrogate models. For the same numerical relativity simulation we find that NRSUR7DQ4 – the leading generic-spin numerical relativity surrogate model[26] – more accurately captures the true parameters of the binary as expected (see Supplementary Figure 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 the NRSUR7DQ4, we find that SEOBNRv5PHM. IMRPHENOMTPHM and IMRPHENOMXPHM can more faithfully describe numerical relativity than NRSUR7DQ4. Specifically, based on mismatches against the `CF_52` simulation [13] – a single spin, mass ratio 1 : 4 simulation with primary dimensionless spin magnitude 0.6 at total masses $75\,M_\odot$, $80\,M_\odot$ and $85\,M_\odot$ – we estimate that according to our NR informed approach, SEOBNRv5PHM would be ≈ 140 times more likely to be used than NRSUR7DQ4 in this region of parameter space due to its improved accuracy. While corner cases such as this exist, NRSUR7DQ4 is still suitable as a proxy for numerical relativity in this work since we are performing Bayesian inference on numerical relativity simulations where the surrogate is the most accurate model [26]. We emphasise that once 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.

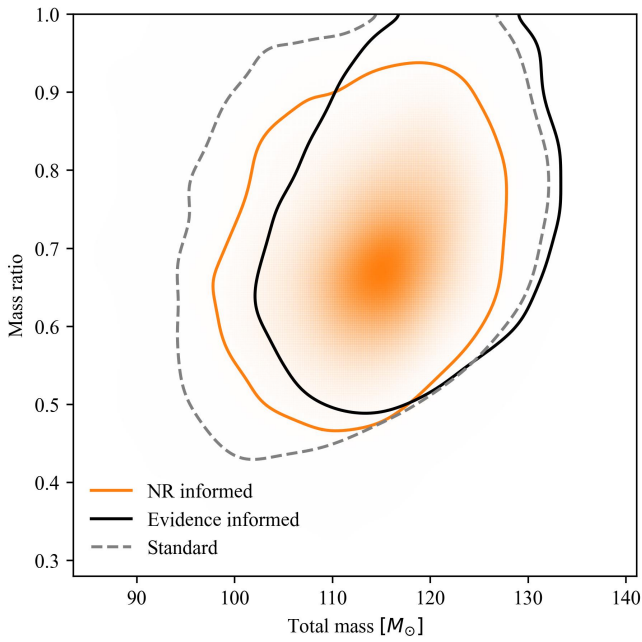Although not presented in this section (see Supplementary Fig-

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

ures 2 and 3), we also analysed the SXS:BBH:0143[15,40] and SXS:BBH:1156[15,40] 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 three methods, with most of the one-dimensional marginalized 90% confidence intervals containing the true value. Our analysis of this case uses 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 the case of SXS:BBH:1156, we find that our method outperforms the Standard and Evidence informed analyses despite partly 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.

Finally, we apply our technique to a real gravitational-wave signal: GW191109_010717 was observed on the 9 November 2019 [29], and has sparked interest within the community since its source likely has large component masses that lie within the upper mass gap – theories suggest that the maximum black hole mass from stellar collapse is $\sim 65\,M_\odot$ [54]. As shown in Figure 3, by incorporating model accuracy in GW Bayesian inference for the first time, we more tightly constrain the total mass of GW191109_010717 to $100 < M < 124\,M_\odot$, and demonstrate that the source of GW191109_010717 has conclusively unequal component masses (assuming the binary black

hole hypothesis). Our re-analysis shows that when using consistent priors and sampler settings as the LIGO–Virgo–KAGRA collaboration, the primary component mass of GW191109_010717 has a 69% probability that it lies within the upper mass gap, consistent with previous work where GW191109_010717 was re-analysed with NR-SUR7DQ4 [55]. This is compared to 51% probability from the LIGO–Virgo–KAGRA analysis [29], thereby significantly increasing the probability that GW191109_010717 was produced from a hierarchical formation mechanism where 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.

3 Conclusions

In this work we present a method to incorporate 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 it (i) marginalizes over model uncertainty by prioritising the most accurate model in each region of 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. Although the approach preferentially uses the most accurate model in each region of parameter space, there is no guarantee that this model is accurate enough to avoid biases in the parameter estimates. However, GW models are 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. Similarly, with more numerical relativity simulations produced, the accuracy of this method increases, 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 in the future.

4 Methods

4.1 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[28],

$$\langle h_m | h_s \rangle = 4\Re \int_{f_{\min}}^{f_{\max}} \frac{\tilde{h}_m^* \tilde{h}_s}{S_n(f)} \, \mathrm{d}f, \tag{1}$$

where a tilde denotes a Fourier transform, * denotes complex conjugation, and $S_n(f)$ is the noise power spectral density, which in this work is Advanced LIGO's design sensitivity[45]. The mismatch[28] between two signals is computed by optimising the normalised inner product over a set of (intrinsic or extrinsic) model parameters $\boldsymbol{\lambda}_m$,

$$\mathcal{M} = 1 - \max_{\boldsymbol{\lambda}_m} \frac{\langle h_m | h_s \rangle}{\sqrt{\langle h_m | h_m \rangle \langle h_s | h_s \rangle}}. \tag{2}$$

The intrinsic parameter space for a generic quasi-circular 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 polarisation $\{\iota, \psi\}$; and overall constant time and phase shifts $\{t_c, \varphi_c\}$ 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 optimise the match over the set of model parameters $\boldsymbol{\lambda}_{\mathrm{m}} = \{t_{\mathrm{c}}, \varphi_{\mathrm{c}}\}$. We maximise over t_{c} via an inverse fast Fourier transform and over ϕ_{c} using the Nelder-Mead optimization algorithm found in SCIPY's `minimize` function[56].

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 polarisability[57] κ. We then prepare an evenly-spaced signal grid over the $\{\kappa, \varphi_{\mathrm{c}}, \iota\}_{\mathrm{s}} \in [0, \pi/2] \otimes [0, 2\pi) \otimes [0, \pi]$ space with $7 \times 6 \times 7 = 294$ elements. At each point in this signal grid, we compute the sky-optimised mismatch[5,57,58] between the signal and the model template from equation (2), where the parameter set we optimise over is $\boldsymbol{\lambda}_{\mathrm{m}} = \{t_{\mathrm{c}}, \varphi_{\mathrm{c}}, \kappa, \varphi_{\mathrm{spin}}\}$. Here φ_{spin} represents the freedom to rotate the in-plane spin (azimuthal) angles ϕ_1, ϕ_2 of $\mathbf{S}_1, \mathbf{S}_2$ by a constant amount. κ is optimised over analytically and optimisations over φ_{c} and φ_{spin} are performed numerically using dual annealing algorithms[3,21,58]. Note that there is no universally agreed upon grid for $\{\kappa, \varphi_{\mathrm{c}}\}$[25,58,59], nor for ι[3,60]. Our specific choice for the $\{\kappa, \varphi_{\mathrm{c}}\}$ grid is based on recent work[5]. Our ι grid spacing is also consistent with the literature[60], but extended to π because the "up-down" symmetry of the GW multipoles with respect to the orbital plane is broken due to precession[61–66].

With these optimisations, we arrive at the maximum possible match between the template and the signal at a given point $\{\kappa, \varphi_{\mathrm{c}}, \iota\}_{\mathrm{s}}$ 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

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

This is done to marginalise over any dependence of the mismatch on the sky position and inclination, thus obtaining 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, $1 - \mathcal{M}_{\mathrm{av}}$, is a discretely averaged version of the sky-and-polarisation averaged faithfulness given by equation (35) of Ramos-Buades *et al.*[2]. For the remainder of this article, we drop the subscript "av" from $\mathcal{M}$.

4.2 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,

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

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[67]. 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,

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

where $p(\mathfrak{M}_i|d)$ 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. 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[33]. 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, 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 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[33].

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[30,31] or in some cases, the model accuracy is averaged over the parameter space of interest[33]. However, a parameter-space dependent prior for the choice of model may solve this problem[33]. For instance, a j-dimensional vector of model parameters can be drawn from the prior and $\Pi(\mathfrak{M}_i|\boldsymbol{\lambda})$ 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[33,39], we use the following model prior conditional on the parameters $\boldsymbol{\lambda}$,

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

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 likely be used to evaluate the likelihood in each region of parameter space.

While several mismatch-dependent priors were tested, equation (6) was chosen for this work since it accentuates small differences in the mismatch between models, and was found to perform optimally. However, since 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 due to small variations on a day-to-day basis from noise artefacts. As a result, it is possible that the relative model probabilities in equation (6) will vary for dif-

ferent power spectral density realisations. We note that this is a common problem in GW astronomy with, *e.g.*, search pipelines similarly using a single representative power spectral density when constructing template banks[68]. We leave a more detailed analysis investigating the choice and stability of this distribution to future work.

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 for each of the GW models used in this analysis.

Owing to computational limitations, we do not have numerical relativity simulations for all possible regions of the compact binary parameter space. For the aligned-spin interpolant construction we therefore evaluate mismatches using the numerical relativity hybrid surrogate model NRHYBSUR3DQ8[27] as a proxy for numerical relativity simulations. There is a long and productive history of GW signal modelling using a variety of approaches[3,21,22,24,58,59,69–84], and we compare against the IMRPHENOMXHM[82] and IMRPHENOMTHM[84] 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[2,24] 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 use the models described in Section 2: IMRPHENOMXPHM[21] (with the updated precession formalisation[25]), IMRPHENOMTPHM[22] and SEOBNRV5PHM[2], and we compare the precessing models against the numerical relativity waveform surrogate model NRSUR7DQ4[26,27] as a proxy for full numerical relativity simulations when computing mismatches.

We next 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. To simplify the construction of the mismatch interpolant for this test application, we reduce the dimensionality of the mismatch parameterisation by artificially fixing several signal and model parameters. We choose to fix the total mass of the binary to $M = 90 M_\odot$ and the inclination angle to $\theta_{\text{JN}} = \pi/3$, where θ_{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 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 to $\log_{10} \mathcal{M}$ of the form

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

with the fit coefficients f_{abc} computed using MATHEMATICA's Fit function and exported to PYTHON using FortranForm. These mismatch surfaces are well-behaved and we find 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 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[46] via Bilby[47]. 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 (6), 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[85] since it is commonly used in gravitational-wave astronomy[86,87]. The Jensen-Shannon Divergence ranges between 0 bits, statistically identical distributions, and 1 bits, statistically distinct distributions. A general rule of thumb is that a Jensen-Shannon Divergence $<$ 50 mbits implies that the distributions are in good agreement[86].

In Supplementary 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 mbits, implying that the distributions are close to statistically identical. We find that the Bayesian analysis that used the interpolant completed in $\sim$ 500 CPU hours, $\sim$ 250× faster than the Bayesian analysis that computed the mismatch exactly. Given the almost statistically identical posteriors and reduced computational cost, we use the interpolated mismatch for all subsequent analyses.

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

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 optimisation over $\boldsymbol{\lambda}_{\text{m}} = \{t_c, \phi_c, \kappa, \varphi_{\text{spin}}\}$ that needs to be repeated for every element of the 294-term sum in equation (3). For example, depending on 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 must keep in mind computational economics when generating the data to construct the fits.

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, equations (2, 3) has been well documented[2,5,22,58,81].

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[88]. 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: $\{\chi_{\mathrm{p}}, \chi_{\mathrm{eff}}\}$. The former is given by[78,89]

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

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

χ_{eff} is the parallel projection counterpart to χ_{p}. It reads[78,90–92]

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

This is a conserved quantity up to 1.5 post-Newtonian (PN) order[91] 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 exist in the literature[60,93], but the one which we empirically determined to be the best for fitting is $\chi_\perp$ given by[94]

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

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 χ_{p}[95], 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 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 χ_{eff}, we introduce

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

We thus have several choices for each perpendicular/parallel scalar: $x = \chi_{\mathrm{p}}$ or $\chi_\perp$, $y = \chi_{\mathrm{eff}}$ or $\chi_\|$, 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 χ_{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: $\{\chi_\perp, \chi_{\mathrm{eff}}\}$ and $\{\chi_\perp, \chi_\|\}$. Accordingly, we introduce the fit training-set labels $K_1 = \{\chi_\perp, \chi_{\mathrm{eff}}, \eta, M\}$ and $K_2 = \{\chi_\perp, \chi_\|, \eta, M\}$.

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[26] of $4300M$ 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[2,5,22,58,81], 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 $\{\chi_\perp, \chi_{\mathrm{eff}}\}$ and $\{\chi_\perp, \chi_\|\}$ 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 $\chi_\perp$-χ_{eff} plane is scatter-plotted as the blue dots in the left panel of Supplementary Figure 4, 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 Supplementary Figure 4). With A and μ fixed, we create an elliptical grid 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 $\{x, y\} = \{\chi_\perp, \chi_{\text{eff}}\}$ in the left and middle panels of Supplementary Figure 4 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 Supplementary Figure 4. 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 kth 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 Supplementary Figure 4, where we place over each red dot a faint blue dot representing the grid points that our algorithm finds. 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 $\{\chi_\perp, \chi_{\text{eff}}\}$ grid of Supplementary Figure 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 obtain, $\sim \mathcal{O}(10^{-2})$, 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 NR-SUR7DQ4'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, therefore, we go one step further and 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

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

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

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

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 fit's extrapolation behaviour 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 $M < 75 M_\odot \cup M > 150 M_\odot$. 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 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 in the case of Supplementary Figure 4, for example, 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 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 $N = N_r \times (N_s + 1)$ 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).

We start our fit optimization routine with $\{n_i, n_j, n_k\} = \{4, 3, 2\}$ and generate fits up to some $\{n_i^{\max}, n_j^{\max}, n_k^{\max}\}$ that ensures that the total number of fit parameters is $\approx N$. The choice of $\{4, 3, 2\}$ yields 160 fit parameters. Smaller values of $\{n_i, n_j, n_k\}$ result in fewer than 100 fitting coefficients and leads to underfitting for training grids of size $\gtrsim \mathcal{O}(200)$ which, as we explain below, is the grid size that we adopt. Our routine picks as the final fit the one for which the values of $\{n_i, n_j, n_k\}$ in equation (20) yield the lowest relative difference with respect to the *training* data set. For this purpose, we define the relative difference between the data and the fit at the kth point

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

to introduce two quantities to gauge fit quality during training. The first is the l^2-norm of Δ_{rel}^k between the fit and the data normalized by the length of the vector

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

and the second is the signed average relative difference

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

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

8

We pick the values for $\{n_i, n_j, n_k\}$ that simultaneously minimize both of the above relative differences. These relative differences are the most important fit attributes for this work as we must robustly predict the mismatches to NR in order to appropriately select which model to use at a given point in the parameter space. If more than one set of values for $\{n_i, n_j, n_k\}$ are returned, we opt for the set which yields a reduced chi square (χ^2/DoF) closest to unity. Once the fit training is complete via the above optimization of $\{n_i, n_j, n_k\}$, we check fit performance over an appropriate verification set which we discuss further below.

The question of training grid resolution can only be answered after setting a target fit unfaithfulness threshold. Here, we aim for $\Delta_{\mathrm{rel}}^{|\mathrm{av}|} \approx 0.05$ for each fit, where

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

is the average absolute relative disagreement between the fit and the verification data. With the above threshold established, we set out to determine whether or not the $\{x, y\}$ training grid of size 10×25 suffices. First, we downsample this grid to create coarser grids of dimension 10×13, 5×13, and 5×7 and compute $\Delta_{\mathrm{rel}}^{|\mathrm{av}|}$ for each with respect to the original verification set (of size 250). As we gradually increase the grid size from 5×7 to 10×25, we observe $\Delta_{\mathrm{rel}}^{|\mathrm{av}|}$ decreasing from ≈ 0.10 to $\lesssim 0.05$ for the fits listed in Supplementary Table 2. For example, the fit used to make Supplementary Figure 5 yields $\Delta_{\mathrm{rel}}^{|\mathrm{av}|} = 0.048$. Increasing the grid size to $\mathcal{O}(1000)$ elements should further reduce $\Delta_{\mathrm{rel}}^{|\mathrm{av}|}$. However, this quickly turns into a problem of diminishing returns given that it would take one month to generate this 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 suffices to capture the dominant trends in the data at the level of fit unfaithfulness that we seek, i.e., $\Delta_{\mathrm{rel}}^{|\mathrm{av}|} \approx 0.05$. Thus, our aforementioned grid of dimensions 10×25 has sufficient resolution. We leave fit improvements to future work which we have already begun undertaking.

In Supplementary Figure 5, we show a contour plot of the unfaithfulness of the fit for the $\log_{10}$ of the NRSUR7DQ4-SEOBNRV5PHM mismatches to the verification data set. Due to the computational burden of obtaining the mismatches, we use data with $y = \chi_{\parallel}$ (χ_{eff}) to verify the data trained with χ_{eff} ($\chi_{\parallel}$). This results in verification sets that are 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 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 $\Delta_{\mathrm{rel}}^{(2)}$ applied to the verification set, but rather the absolute value of the summand in equation (23). The fits for the NRSUR7DQ4-IMRPHENOMTPHM and the NRSUR7DQ4-IMRPHENOMXPHM mismatches also yield similar level of agreement as do the fits trained by the $y = \chi_{\mathrm{eff}}$ set.

We summarize these results and provide additional metrics for all the fits in Supplementary Table 2 where we see that the average relative distance (22) between each fit and the corresponding verification data is always less than 4×10^{-3} and the average relative difference (23) has magnitude less than 0.02. Interestingly, we observe that $\Delta_{\mathrm{rel}}^{(2),\mathrm{ver}}$ 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 χ^2/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 $M < 75 M_{\odot} \cup M > 150 M_{\odot}$. Recall that we chose the particular form of equation (20) for the fitting function to better control unwanted extrapolation 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 Supplementary Figure 6, 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 the $\{\eta, \bar{S}_{1,2}\} = \{0.122, 0.9\}$ ($q = 1/6$) 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 Supplementary Figure 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 $M = 50 M_{\odot}$ to $200 M_{\odot}$, 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 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 $q \geq 1/5$, $\bar{S}_i \leq 0.85$.

As a further test of the fit's performance in its extrapolation regime, we performed two more parameter recoveries of an NR simulation, SXS:BBH:1156, with the injected value for M set to $75 M_{\odot}$ and $100 M_{\odot}$, and $q = 0.228$ placing the former "squarely" outside the $M > 75 M_{\odot}, q \geq 1/4$ training regime of our fits. This simulation also has $|\mathbf{S}_{2,\perp}|/m_2^2 \approx 0.76$ while the primary has negligible planar spin. We show the results of our method applied to this Bayesian inference analysis in Supplementary Figure 3 with the 2D posteriors from the $M = 75 M_{\odot}$ ($100 M_{\odot}$) injections plotted in the top (bottom) panels. We can see that in both analyses, our method recovers the injected values for m_1, m_2 and the effective spins very well with the majority of the samples clustered near the injected values. Such a good recovery of the masses, hence the mass ratio, for the $M = 75 M_{\odot}$ injection is an indirect testament to the robustness of the fit (20), especially since slightly more than half the total mass posteriors for this particular Bayesian analysis happen to be less than $75 M_{\odot}$.

We conclude this section by briefly returning to two issues: the first regarding the fact that the fitting variables $\{\eta, x, y\}$ used to construct the fit are not fully independent of each other, as each one is a function of the mass ratio q, and the second regarding the choice of power spectral density used when calculating the mismatch and hence the fits. For the first issue, 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 $K_1(y = \chi_{\mathrm{eff}})$, $K_2(y = \chi_{\parallel})$. 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 Supplementary Table 2.

For the second issue, as already explained, we use the Advanced LIGO's design sensitivity[45] when computing the mismatch and this will subtly change when the profile of the power spectral density is varied. If this method were to be used during live observing run periods, where the power spectral density is changing on a day-to-day basis 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 [68]. The fit would then be reconstructed prior to each GW observing run.

Data Availability

The aligned- and generic-spin match interpolants as well as the posterior samples from the analyses performed in this work are available in our data release (Hoy et al., in prep.).

Code Availability

Python scripts detailing our modifications to `Bilby`[47] are available in our data release (Hoy et al., in prep.).

Acknowledgements

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`[47] and performed all parameter estimation analyses. JT generated the aligned-spin mis-

matches and the aligned-spin interpolant. SA constructed the generic-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. All authors contributed to the interpretation of the results and wrote the paper.

Competing interests

The authors declare no competing interests.

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. *Phys. Rev. D* **110**, 064024 (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. Dhani, A. *et al.* Systematic Biases in Estimating the Properties of Black Holes Due to Inaccurate Gravitational-Wave Models. *arXiv e-prints* (2024). 2404.05811.

7. Pürrer, M. & Haster, C.-J. Gravitational waveform accuracy requirements for future ground-based detectors. *Phys. Rev. Res.* **2**, 023151 (2020). 1912.10055.

8. 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.

9. 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.

10. Aasi, J. *et al.* Advanced LIGO. *Class. Quant. Grav.* **32**, 074001 (2015). 1411.4547.

11. Acernese, F. *et al.* Advanced virgo: a second-generation interferometric gravitational wave detector. *Classical and Quantum Gravity* **32**, 024001 (2014).

12. Akutsu, T. *et al.* Overview of KAGRA: Detector design and construction history. *PTEP* **2021**, 05A101 (2021). 2005.05574.

13. Hamilton, E. *et al.* Catalog of precessing black-hole-binary numerical-relativity simulations. *Phys. Rev. D* **109**, 044032 (2024). 2303.05419.

14. 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.

15. Boyle, M. *et al.* The SXS Collaboration catalog of binary black hole simulations. *Class. Quant. Grav.* **36**, 195006 (2019). 1904.04831.

16. Healy, J., Lousto, C. O., Zlochower, Y. & Campanelli, M. The RIT binary black hole simulations catalog. *Class. Quant. Grav.* **34**, 224001 (2017). 1703.03423.

17. 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.

18. Healy, J. & Lousto, C. O. Third RIT binary black hole simulations catalog. *Phys. Rev. D* **102**, 104018 (2020). 2007.07910.

19. Healy, J. & Lousto, C. O. Fourth RIT binary black hole simulations catalog: Extension to eccentric orbits. *Phys. Rev. D* **105**, 124010 (2022). 2202.00018.

20. Jani, K. *et al.* Georgia Tech Catalog of Gravitational Waveforms. *Class. Quant. Grav.* **33**, 204001 (2016). 1605.03204.

21. 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.

22. 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.

23. 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). 2111.14149.

24. 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.

25. 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. *arXiv e-prints* (2024). 2412.16721.

26. Varma, V. *et al.* Surrogate models for precessing binary black hole simulations with unequal masses. *Phys. Rev. Research.* **1**, 033015 (2019). 1905.09300.

27. Varma, V. *et al.* Surrogate model of hybridized numerical relativity binary black hole waveforms. *Phys. Rev. D* **99**, 064045 (2019). 1812.07865.

28. Owen, B. J. Search templates for gravitational waves from inspiraling binaries: Choice of template spacing. *Phys. Rev. D* **53**, 6749–6761 (1996). gr-qc/9511032.

29. 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). 2111.03606.

30. Ashton, G. & Khan, S. Multiwaveform inference of gravitational waves. *Phys. Rev. D* **101**, 064037 (2020). 1910.09138.

31. 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.

32. Ashton, G. & Dietrich, T. The use of hypermodels to understand binary neutron star collisions. *Nature Astron.* **6**, 961–967 (2022). 2111.09214.

33. Hoy, C. Accelerating multimodel Bayesian inference, model selection, and systematic studies for gravitational wave astronomy. *Phys. Rev. D* **106**, 083003 (2022). 2208.00106.

34. 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.

35. Doctor, Z., Farr, B., Holz, D. E. & Pürrer, M. Statistical Gravitational Waveform Models: What to Simulate Next? *Phys. Rev. D* **96**, 123011 (2017). 1706.05408.

36. 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). 1903.09204.

37. Read, J. S. Waveform uncertainty quantification and interpretation for gravitational-wave astronomy. *Class. Quant. Grav.* **40**, 135002 (2023). 2301.06630.

38. Khan, S. Probabilistic model for the gravitational wave signal from merging black holes. *Phys. Rev. D* **109**, 104045 (2024). 2403.11534.

39. Jan, A. *Marginalizing over the waveform systematics of compact binary coalescence models using RIFT*. Master's thesis, Rochester Institute of Technology (2021).

40. Blackman, J. *et al.* Numerical relativity waveform surrogate model for generically precessing binary black hole mergers. *Phys. Rev. D* **96**, 024058 (2017). 1705.07089.

41. 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).

42. 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). 1908.05707.

43. 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). 2408.03410.

44. 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).

45. 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.

46. 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.

47. Ashton, G. *et al.* BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy. *Astrophys. J. Suppl.* **241**, 27 (2019). 1811.02042.

48. 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.

49. Ajith, P. *et al.* Inspiral-merger-ringdown waveforms for black-hole binaries with non-precessing spins. *Phys. Rev. Lett.* **106**, 241101 (2011). 0909.2867.

50. 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.

51. 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.

52. Toubiana, A. & Gair, J. R. Indistinguishability criterion and estimating the presence of biases. *arXiv e-prints* (2024). 2401.06845.

53. Hannam, M. *et al.* General-relativistic precession in a black-hole binary. *Nature* **610**, 652–655 (2022). 2112.11300.

54. Woosley, S. E., Heger, A. & Weaver, T. A. The evolution and explosion of massive stars. *Rev. Mod. Phys.* **74**, 1015–1071 (2002).

55. Islam, T. *et al.* Analysis of GWTC-3 with fully precessing numerical relativity surrogate models. *arXiv e-prints* (2023). 2309.14473.

56. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**, 261–272 (2020).

57. 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.

58. 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.

59. 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.

60. 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.

61. Bruegmann, B., Gonzalez, J. A., Hannam, M., Husa, S. & Sperhake, U. Exploring black hole superkicks. *Phys. Rev. D* **77**, 124047 (2008). 0707.0135.

62. 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). 0902.4077.

63. Kalaghatgi, C. & Hannam, M. Investigating the effect of in-plane spin directions for precessing binary black hole systems. *Phys. Rev. D* **103**, 024024 (2021). 2008.09957.

64. 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). 2001.10936.

65. 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). 2310.16980.

66. 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). 2402.00813.

67. Skilling, J. Nested sampling for general bayesian computation. *Bayesian Anal.* **1**, 833–859 (2006). URL https://doi.org/10.1214/06-BA127.

68. Dal Canton, T. & Harry, I. W. Designing a template bank to observe compact binary coalescences in Advanced LIGO's second observing run. *arXiv e-prints* (2017). 1705.01845.

69. 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.

70. 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.

71. 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.

72. 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.

73. 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.

74. 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.

75. 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`.

76. 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`.

77. 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`.

78. Hannam, M. *et al.* Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms. *Phys. Rev. Lett.* **113**, 151101 (2014). `1308.3271`.

79. 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`.

80. 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). `1911.06050`.

81. 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`.

82. 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`.

83. 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`.

84. 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`.

85. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Info. Theor.* **37**, 145–151 (1991).

86. 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`.

87. 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`.

88. Schmidt, P., Harry, I. W. & Pfeiffer, H. P. Numerical Relativity Injection Infrastructure. *arXiv e-prints* (2017). `1703.01076`.

89. 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`.

90. Damour, T. Coalescence of two spinning black holes: An effective one-body approach. *Phys. Rev.* **D64**, 124013 (2001). `gr-qc/0103018`.

91. Racine, E. Analysis of spin precession in binary black hole systems including quadrupole-monopole interaction. *Phys. Rev.* **D78**, 044021 (2008). `0803.1820`.

92. 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`.

93. 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`.

94. 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`.

95. Gerosa, D. *et al.* A generalized precession parameter χ_{P} to interpret gravitational-wave data. *Phys. Rev. D* **103**, 064067 (2021). `2011.11948`.

Supplementary Material

To demonstrate the improved accuracy of the NR informed approach compared to Evidence informed and Standard algorithms, we compared the posterior distributions obtained with each method against those obtained with NRSUR7DQ4; the leading generic spin numerical relativity surrogate model. We only considered the SXS:BBH:0926 numerical relativity simulation since it was the main focus of this work.

In Supplementary Figure 1 we see that our NR informed approach infers a posterior that significantly overlaps with NRSUR7DQ4. For the inference of the binary masses we see that NR informed has support for slightly larger primary masses, but the secondary mass remains comparable with NRSUR7DQ4. Similarly, when considering the inferred component spins, we see that our NR informed approach infers a comparable marginalized one-dimensional posterior for the effective perpendicular spin, with slightly more support for larger effective parallel spin. Importantly, we see that the Evidence informed and Standard approaches fail to capture the true value. Specifically, the Evidence informed result has low overlap with the NRSUR7DQ4 result and the Standard result unnecessarily inflates the uncertainty for the effective perpendicular spin. Interestingly, while both our NR informed approach and NRSUR7DQ4 capture the true effective spins of the binary with their two-dimensional 90% credible intervals, the majority of posterior support is to lower effective perpendicular and larger effective parallel spins. In general, we found that our NR informed approach obtained the most statistically similar one-dimensional posterior probability distributions to the surrogate out of the methods considered in this work.

We additionally analysed the SXS:BBH:0143[15,40] and SXS:BBH:1156[15,40] numerical relativity simulations produced by the SXS collaboration with our NR informed approach, as well as the Evidence informed and Standard algorithms. For the case of SXS:BBH:0143, we found largely overlapping posteriors between methods, with most of the one-dimensional marginalized 90% confidence intervals containing the true value. For the case of SXS:BBH:1156, we found that our method outperforms the Standard and Evidence informed analyses despite being in the extrapolation regime of our technique. Here, we provide further details.

In Supplementary Figure 2 we show the results obtained with our method, NR informed, vs. the Evidence informed and Standard analyses for the parameter recovery of the SXS:BBH:0143 numerical relativity simulation. We see that in general, all methods give largely overlapping posteriors. When inspecting the one-dimensional marginalized posterior distributions, provided in the associated data release, the biggest disagreement is seen in the inferred secondary spin magnitude of the binary: our method prefers a rapidly spinning black hole, while the other methods are largely uninformative with a spin ~ 0.25. The reason is because our method primarily chooses 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[48], it is not surprising that the biggest difference between methods is seen for the secondary spin. Since the secondary spin contribution is also suppressed for asymmetric mass ratio binaries, it is possible that the mismatch interpolant is missing information, leading to poorly chosen models in this region of the parameter space. A mismatch interpolant that is a function of the individual spin components may solve this problem, but it would require a significantly larger mismatch dataset. We leave this to future work. It could also be independent of the interpolant: the interpolant could be accurately describing the mismatch, but SEOBNRv5PHM may not be accurate enough to avoid parameter biases, despite being the most accurate model of those considered in this region of the parameter space.

In Supplementary Figure 3 we compare results for the SXS:BBH:1156 numerical relativity simulation. Although all methods infer the injected value within 90% credible interval, we see that our method outperforms the others, especially seen for the high total mass ($M = 100\,M_{\odot}$) injection: only our method contains the injected spin within the 40% confidence interval. When inspecting the one-dimensional marginalized posterior distributions, provided in the associated data release, we consistently see a more accurate inference
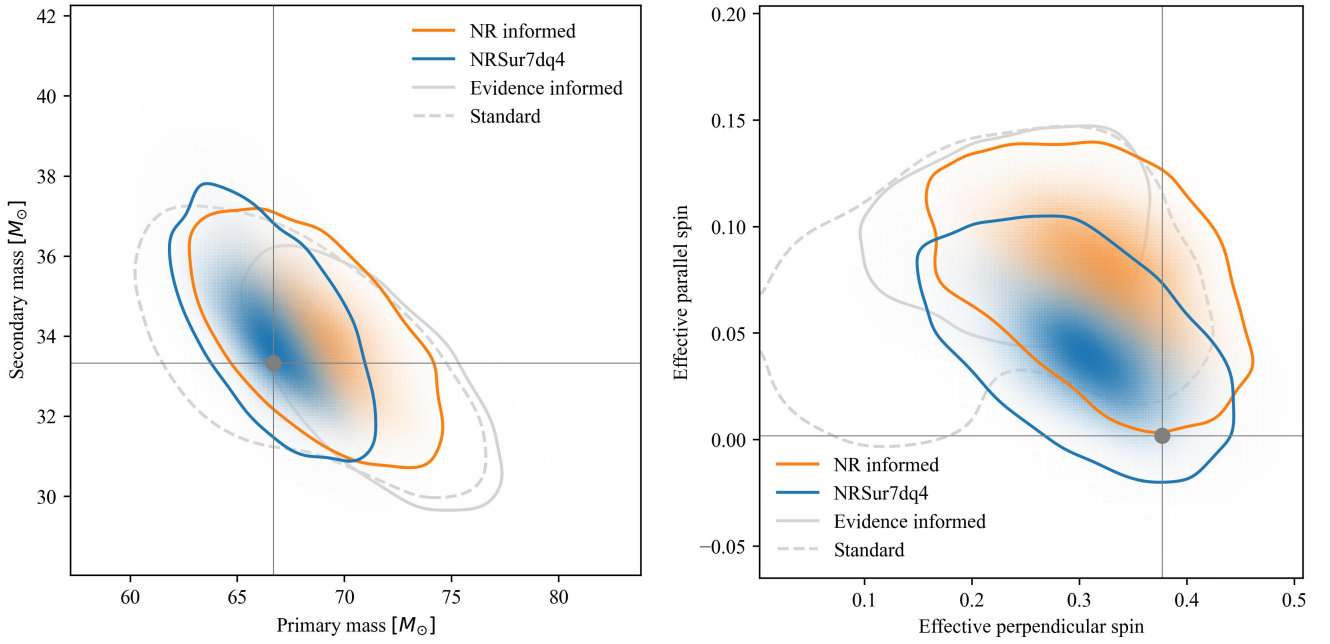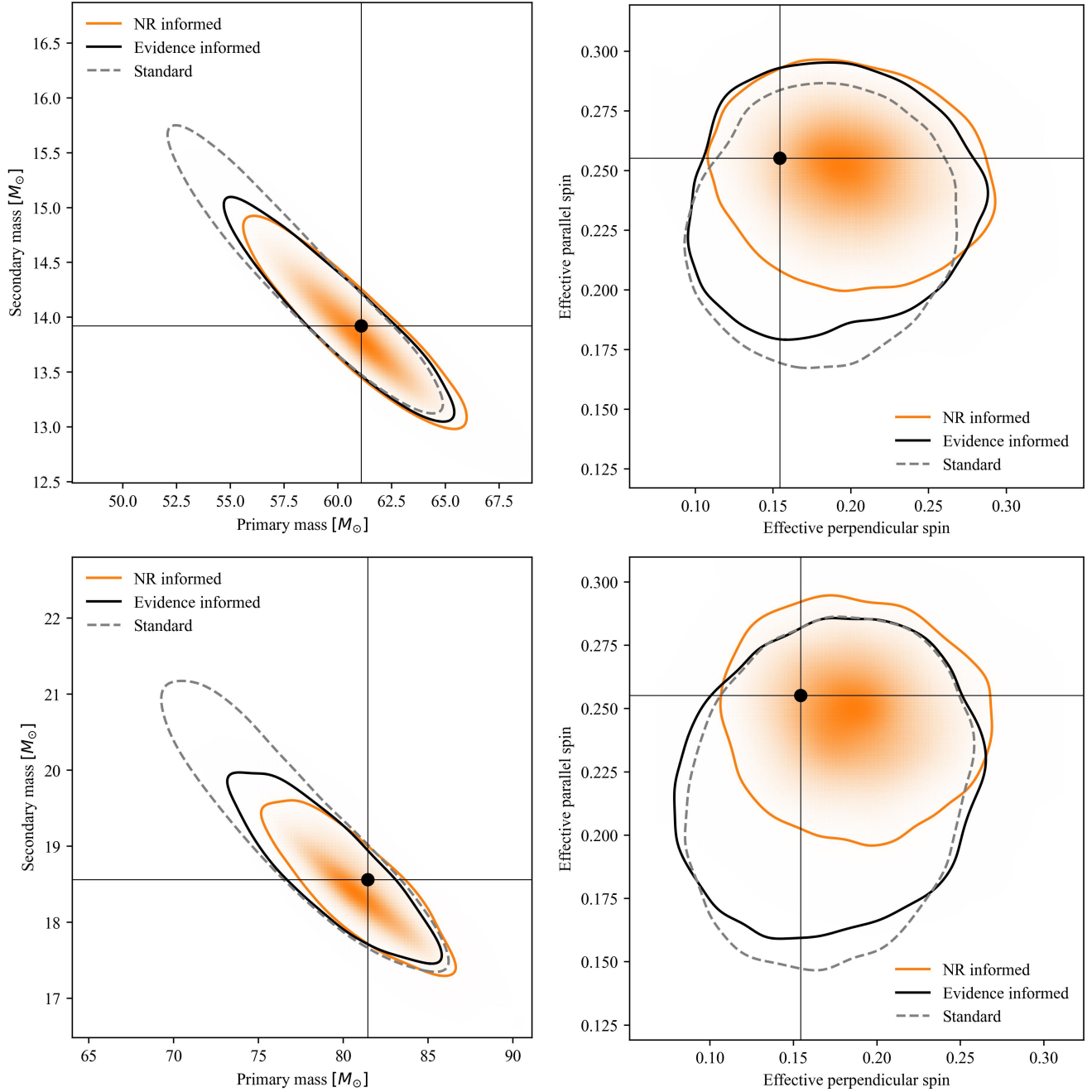
Supplementary Figure 1 | Two-dimensional posterior probabilities obtained in our analysis of the `SXS:BBH:0926` numerical relativity simulation. The left panel shows the measurement of the primary and secondary mass of the binary, and the right panel shows the inferred effective parallel and perpendicular spin components (as defined in equation 11, 12). The contours represent 90% credible intervals and the grey cross hairs indicate the true value. The Evidence-informed and Standard analyses are set to a lighter color than in other Figures to highlight the comparison between NR informed and NRSUR7DQ4.
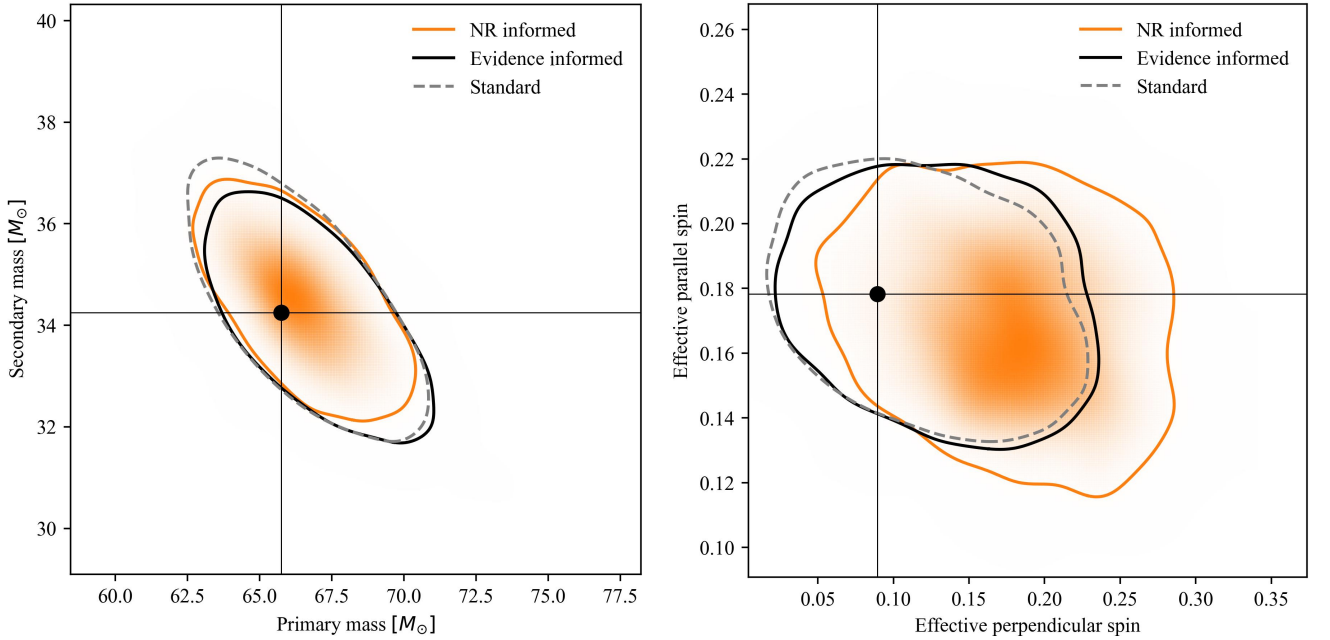
Supplementary Figure 2 | Two-dimensional posterior probabilities obtained in our analysis of the `SXS:BBH:0143` numerical relativity simulation. The left panel shows the measurement of the primary and secondary mass of the binary, and the right panel shows the inferred effective parallel and perpendicular spin components (as defined in equation 11, 12). The contours represent 90% credible intervals and the black cross hairs indicate the true value.

of the binary parameters with our method. As with `SXS:BBH:0143`, we see the worst performance for the inferred spin of the smaller black hole in the binary. Although the injected value is contained within the 90% confidence interval, the Evidence informed and Standard analyses capture the true value more accurately. For the low total mass ($M = 75\,M_\odot$) injection, we observe a better performance by all methods. We argue that this is because the GW produced from a lower mass system spends more time within the sensitive region of the GW detectors, meaning that all GW models better capture the true source parameters of the binary, as long as the primary is not strongly processing.

Supplementary Figure 3 | Supplementary Figure 3: Two-dimensional posterior probabilities obtained in our analysis of the `SXS:BBH:1156` numerical relativity simulation. The left column shows the measurement of the primary and secondary mass of the binary, and the right column shows the inferred effective parallel and perpendicular spin components (as defined in equations (11, 12)). The top row shows a low total mass injection of $M = 75\ M_\odot$ and the bottom row shows a high total mass injection of $M = 100\ M_\odot$. The contours represent 90% credible intervals and the black cross hairs indicate the true value.

|		
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

Supplementary Table 1 | Comparison of Jensen-Shannon divergences when using the true mismatch. Jensen-Shannon divergences between posteriors obtained when using true mismatch and the mismatch interpolant for an aligned-spin injection and recovery. Column one lists the parameters for which we display the Jensen-Shannon divergences of the posteriors in column two. These Jensen-Shannon divergences are reported in the base 2 logarithm and reported in millibits (mbits).
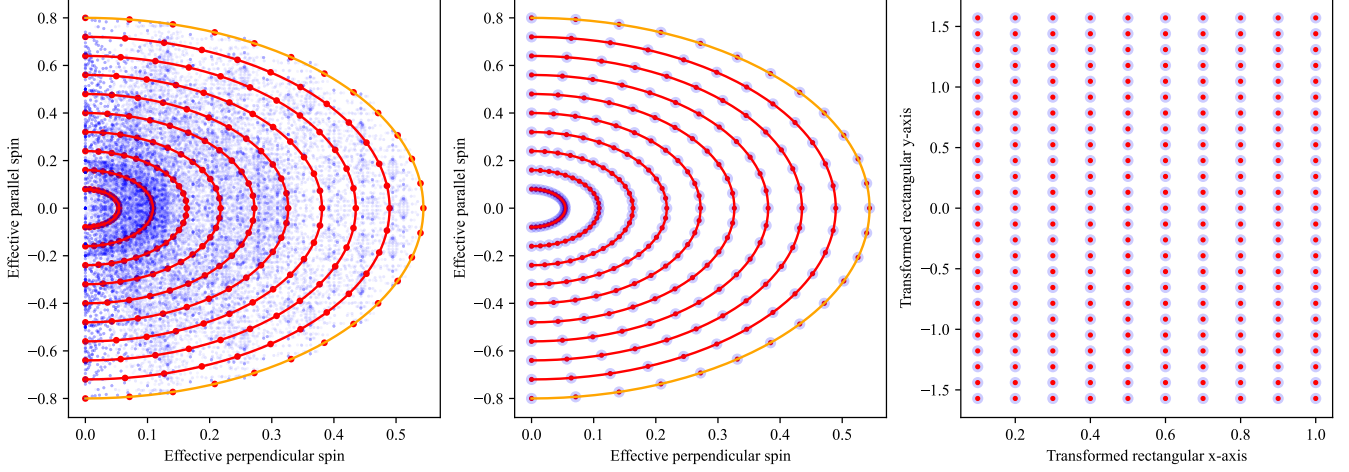
Supplementary Figure 4 | Construction of the two dimensional elliptical and rectangular grids. We show the construction of the two dimensional elliptical and rectangular grids for the $\{x, y\} = \{$Effective perpendicular spin, $\chi_\perp$, Effective spin, $\chi_{\text{eff}}\}$ pairing. The blue dots in the left panel represent the ~ 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 equation (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 $\pi/2$. Though this grid size is somewhat arbitrary, it is limited by computational resources such as the runtimes to generate the mismatches (see Sec. 4.5) and to construct the fits. 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 transformed rectangular coordinates $X = 0.1, 0.2, \ldots, 1, Y = -\pi/2, -\pi/2 + \pi/24, \ldots, \pi/2$ and the blue dots are mapped from the blue dots of the middle panel via equations (19a, 19b).

Model	y	$\{n_i, n_j, n_k\}$	# Params	$\Delta_{\text{rel}}^{(1),\text{ver}}$	$\Delta_{\text{rel}}^{(2),\text{ver}}$	$1 - \tilde{R}^2$	χ^2/DoF
SEOBNRv5PHM	χ_{eff}	$\{4, 3, 2\}$	220	0.0032	-0.013	0.0049	0.050
SEOBNRv5PHM	$\chi_\|$	$\{5, 3, 2\}$	264	0.0026	0.0041	0.0052	1.3
IMRPHENOMTPHM	χ_{eff}	$\{5, 3, 2\}$	264	0.0039	-0.012	0.012	0.99
IMRPHENOMTPHM	$\chi_\|$	$\{4, 3, 2\}$	220	0.0039	-0.0090	0.0077	0.076
IMRPHENOMXPHM-ST	χ_{eff}	$\{4, 3, 2\}$	220	0.0038	-0.019	0.010	1.2
IMRPHENOMXPHM-ST	$\chi_\|$	$\{4, 3, 2\}$	220	0.0032	-0.0021	0.0078	0.74

Supplementary Table 2 | Metrics for the mismatch fitting function. 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 $\chi_\|$ when constructing the elliptical fit training grid, *e.g.*, shown for χ_{eff} in Supplementary Figure 4. 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 (22, 23) 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 $y = \chi_\|$ grid are slightly more faithful to the verification data and $\tilde{R}^2 \gtrsim 0.99$ for all fits.
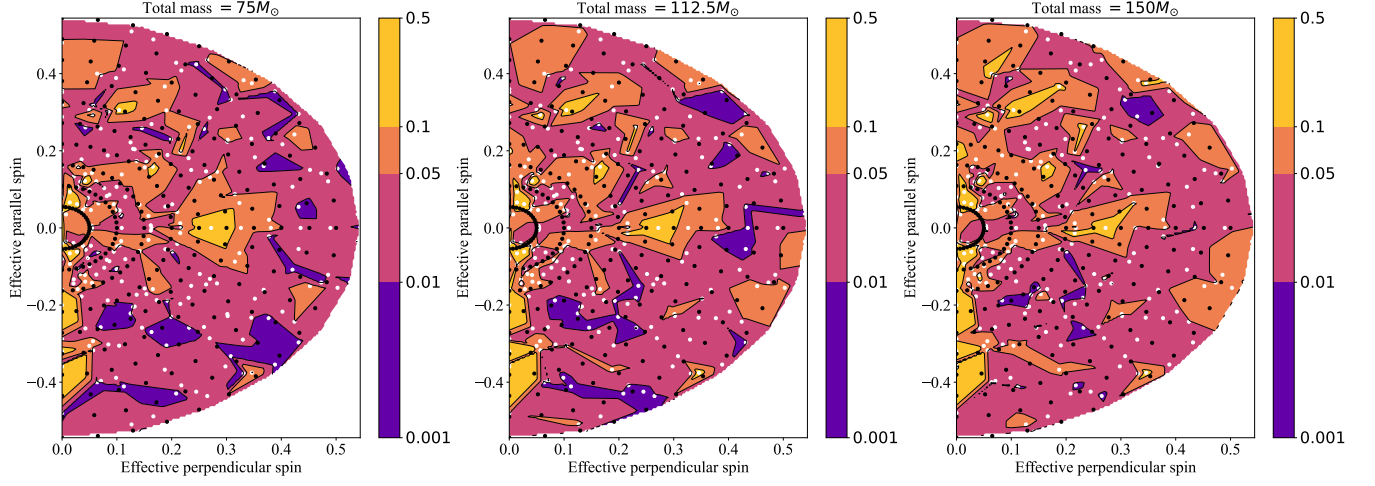
Supplementary Figure 5 | Fit unfaithfulness over the verification set. We plot the absolute value of the relative difference, *i.e.*, $\Delta_{\rm rel}^k$ [see equation (21)], between the data for the $\log_{10}$ of the NRSUR7DQ4-SEOBNRV5PHM mismatches and the fit (20) to it with equations (19a, 19b) substituted. The panels from left to right correspond to total masses of $75M_\odot$, $112.5M_\odot$, $150M_\odot$. The effective perpendicular, $\chi_\perp$, and effective parallel spins, $\chi_\parallel$, 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.
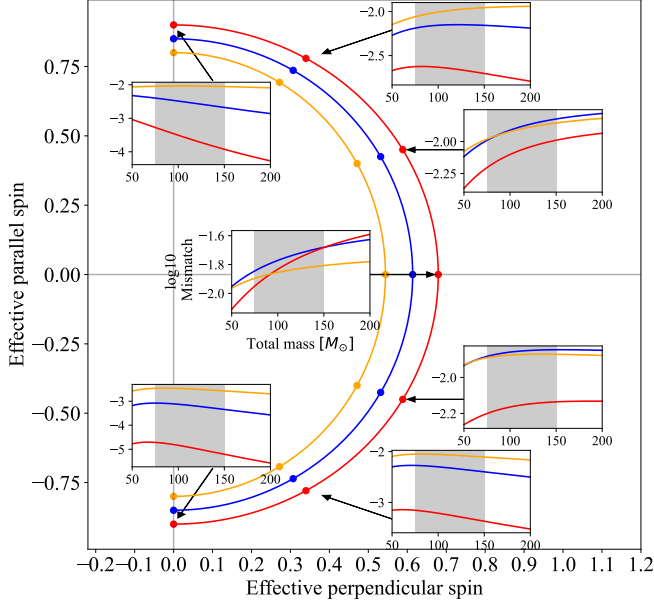
Supplementary Figure 6 | Extrapolation of the fits. We show the performance of the fit to the $\log_{10}$ of the NRSUR7DQ4-SEOBNRV5PHM mismatches extrapolated to $\{q, \bar{S}_{1,2}\} = \{1/5, 0.85\}$ (blue) and $\{1/6, 0.9\}$ (red) as well as the fit at the edge of its training region with $\{q, \bar{S}_{1,2}\} = \{1/4, 0.8\}$ (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 Supplementary Figure 4. Pointing to each $\Phi = $ constant dot is an inset showing the plot of the fit from $M = 50 M_\odot$ to $200 M_\odot$, 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.