

Gravitational waves carry information beyond effective spin parameters but it is hard to extract

Simona J. Miller^{1,2,*}, Zoe Ko^{3,†}, Tom Callister^{4,‡} and Katerina Chatziioannou^{1,2,§}

¹*Department of Physics, California Institute of Technology, Pasadena, California 91125, USA*

²*LIGO Laboratory, California Institute of Technology, Pasadena, California 91125, USA*

³*Department of Physics, University of California Berkeley, Berkeley, California 94720, USA*

⁴*Kavli Institute for Cosmological Physics, The University of Chicago, Chicago, Illinois 60637, USA*



(Received 12 January 2024; accepted 13 March 2024; published 10 May 2024)

Gravitational wave observations of binary black hole mergers probe their astrophysical origins via the binary spin, namely the spin magnitudes and directions of each component black hole, together described by six degrees of freedom. However, the emitted signals primarily depend on two effective spin parameters that condense the spin degrees of freedom to those parallel and those perpendicular to the orbital plane. Given this reduction in dimensionality between the physically relevant problem and what is typically measurable, we revisit the question of whether information about the component spin magnitudes and directions can successfully be recovered via gravitational-wave observations, or if we simply extrapolate information about the distributions of effective spin parameters. To this end, we simulate three astrophysical populations with the same underlying effective-spin distribution but different spin magnitude and tilt distributions, on which we conduct full individual-event and population-level parameter estimation. We find that parametrized population models can indeed qualitatively distinguish between populations with different spin magnitude and tilt distributions at current sensitivity. However, it remains challenging to either accurately recover the *true* distribution or to diagnose biases due to model misspecification. We attribute the former to practical challenges of dealing with high-dimensional posterior distributions, and the latter to the fact that each individual event carries very little information about the full six spin degrees of freedom.

DOI: [10.1103/PhysRevD.109.104036](https://doi.org/10.1103/PhysRevD.109.104036)

I. INTRODUCTION

The spins of black holes (BHs) in binaries (BBHs) are a unique probe of physics on multiple scales, from fundamental BH properties to stellar interiors and the astrophysical environments in which compact binaries form. Each binary possesses six spin degrees of freedom: the spin magnitudes, polar angles (tilts), and azimuthal angles of each binary component [1]. BH spins are encoded in the gravitational waves (GWs) the binary emits and can, at least in principle, be constrained from observation [2,3] by the LIGO [4] and Virgo [5] detectors. The magnitudes and directions of the spins at merger are determined by the spin each BH has upon formation as well as the binary's evolutionary history, e.g. [6–8]. Spin measurements are therefore a promising way to determine whether BBHs form dynamically or in the field, e.g. [9–14], and answer

questions such as the role of angular momentum transfer in stars, tidal interactions, and mass transfer, e.g. [15–19].

Despite their astrophysical importance, spins remain poorly constrained in GW data. Their imprint on the signal is typically subdominant to other intrinsic effects such as the BH masses, e.g. [20–24]. Furthermore, not all six spin degrees of freedom affect the signal equally. Though waveform models formally depend on the full spin vectors [25–28], analytical post-Newtonian calculations indicate that the dominant spin effect is captured by two effective parameters: the effective aligned spin χ_{eff} that includes the spin components parallel to the Newtonian orbital angular momentum [29] and the effective precessing parameter χ_p that includes the perpendicular components [30]. The former primarily affects the length of the signal while the latter describes spin precession, the change in binary orientation due to spin-orbit and spin-spin interactions [31]. Unsurprisingly, then, constraints on the astrophysical distributions of χ_{eff} and χ_p can be typically obtained with fewer observations and are less prone to population model systematics than the spin components [32–34].

Although less well measurable, it is instead the underlying spin components that are of prime astrophysical

*smiller@caltech.edu

†zko@berkeley.edu

‡tcallister@uchicago.edu

§kchatziioannou@caltech.edu

interest. GW signals contain *some* information about component spins. However, unlike χ_{eff} and χ_{p} that appear prominently in the GW phase and amplitude and whose measurability can be predicted with analytic arguments [35,36], individual spin components have a significantly subdominant effect on the waveform. The resulting constraints on the *astrophysical* distribution of spin components are correspondingly weaker and in many cases subject to uncertainties about the role of population models [34,37]. Indeed, even though it is widely accepted that BBHs have a range of χ_{eff} values that are not symmetric about zero [2,3,38] and that not all BBHs have a vanishing χ_{p} [2,3], the exact shape of the inferred distribution for spin magnitudes and directions depends on the parametrization of the corresponding population model. For example, different parametrizations for the angle between the spins and the Newtonian orbital angular momentum lead to varied conclusions about where the distribution peaks and the degree of spin-orbit misalignment [3,34,37,39–41].

Central to this discussion are the questions of how much information GW signals actually contain about the BH spin components versus χ_{eff} alone, how feasible it practically is to reliably extract this information, and the extent to which conclusions are driven by informative data or simply by overly restrictive models. In this paper, we approach these issues by posing three questions, from which we draw conclusions:

- (1) Do GWs carry information about spin components, or are we just extrapolating the effective aligned spin χ_{eff} ? (Sec. IV) *Yes, we can distinguish between populations with low, moderate, and high spins even when they have identical effective spin distributions.*
- (2) Can component spin distributions be accurately measured? (Sec. V) *Even though we can qualitatively tell apart BH populations with different spin distributions, characterizing them accurately is practically challenging.*
- (3) Can we tell when measurements of component spin distributions are biased? (Sec. VI) *Common tests based on posterior predictive checks cannot identify modeling biases in component spin distributions due to the fact that individual-event posteriors are extremely weakly informative about spin components.*

The remainder of this paper presents our analysis in support of these conclusions. We discuss spin degrees of freedom, effective spin parameters, and the notation used throughout in Sec. II. Our methods are briefly described in Sec. III, and are expanded upon in the Appendices. Results about measuring the component spin distributions are presented in Sec. IV. Section V introduces the extensive series of verification methods—both population and individual-event level—we use to ensure the robustness of our results, all of which are further elaborated upon in

the Appendices. In Sec. VI, we identify limitations of the traditional method of using posterior predictive checks to assess biased population measurements, and identify the sources of this bias. We compare our findings to those of past work in Sec. VII, and then conclude in Sec. VIII.

II. SPIN MAGNITUDES AND TILTS VERSUS EFFECTIVE SPIN PARAMETERS

Each BH in the binary is described by a dimensionless spin vector $\vec{\chi}_i$, $i \in \{1, 2\}$. In a coordinate system where the z axis is aligned with the binary Newtonian orbital angular momentum \vec{L} , the spin vector is characterized by a magnitude $\chi_i \in [0, 1]$, polar angle $\theta_i \in [0, \pi]$, and azimuthal angle $\phi_i \in [0, 2\pi]$. Modulo horizon absorption effects, the spin magnitude is constant throughout the binary evolution [42,43], while the spin angles evolve due to spin-orbit and spin-spin interactions causing the spin vector to precess [31,44].

The full six spin degrees of freedom remain relatively poorly constrained by GW signals. Rather, the dominant spin effects are expressed by two effective parameters. The mass-weighted average spin projected onto \vec{L}

$$\chi_{\text{eff}} = \frac{\chi_1 \cos \theta_1 + q \chi_2 \cos \theta_2}{1 + q} \in (-1, 1), \quad (1)$$

is referred to as *effective aligned spin*, where we have defined the binary mass ratio $q \equiv m_2/m_1$, where $m_1 \geq m_2$ are the BH masses. The effective aligned spin is, in general, better constrained as it is related (in the equal mass limit) to the leading-order spin contribution in the post-Newtonian expansion for the GW inspiral phase [1]. Additionally, χ_{eff} is conserved under spin-precession and radiation reaction to at least the second post-Newtonian order [29].

Spin-precession effects are captured with the *effective precessing parameter*

$$\chi_{\text{p}} = \max \left[\chi_1 \sin \theta_1, \left(\frac{3 + 4q}{4 + 3q} \right) q \chi_2 \sin \theta_2 \right] \in [0, 1]. \quad (2)$$

This parameter and its extensions [45,46] are motivated by the fact that spin-orbit precession (and the GW amplitude and phase modulations it induces) are driven by in-plane spin components [30,47]. Constraints on χ_{p} are typically much weaker than χ_{eff} especially given the observed absence of large spin precession in BBHs [2,3]. In what follows, we therefore focus on χ_{eff} .

III. METHODOLOGY

In order to isolate the amount of information included in GW signals about the spin components relative to the effective spin, we simulate astrophysical populations with identical χ_{eff} distributions but different underlying

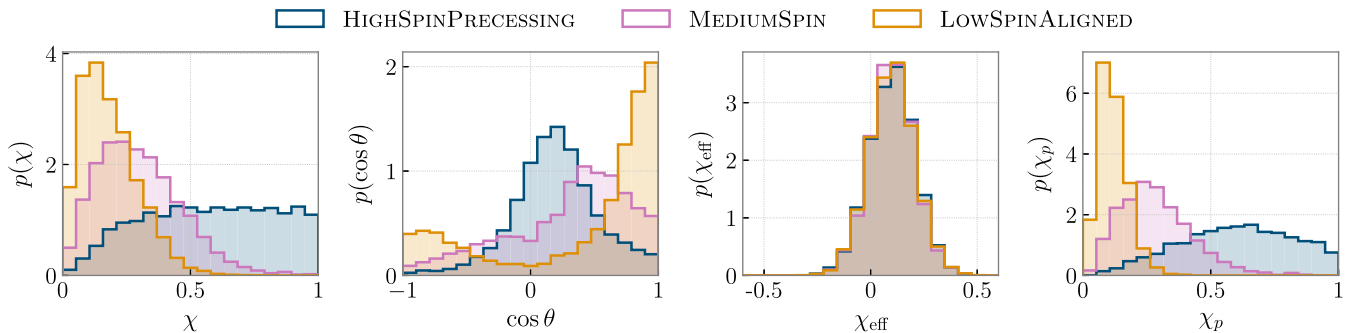


FIG. 1. Spin distributions for the three simulated BBH populations we use to assess the amount of recoverable information GW signals contain about component spin distributions. The three populations share the same χ_{eff} distribution (chosen for consistency with current data), but differ in their underlying component spin and χ_p distributions. From left to right, panels show the spin magnitude χ , spin tilt $\cos \theta$, the effective spin parameter χ_{eff} , and the effective precession parameter χ_p . Navy: HighSpinPrecessing: a BBH population with few vanishing spins that are preferentially oriented close to the orbital plane. Pink: MediumSpin: a BBH population with moderate spin magnitudes peaking at $\chi = 0.25$ and preferentially aligned with the Newtonian orbital angular momentum. Orange: LowSpinAligned: a BBH population with low spin magnitudes peaking at $\chi = 0.10$ with both strongly aligned and antialigned subpopulations.

component spin distributions. We choose a χ_{eff} distribution that is qualitatively similar to current constraints [3,34,48] and decompose it into three populations with distinct spin magnitudes and tilt angle distributions. The azimuthal angles are uniformly distributed. These distributions are not astrophysically motivated, but rather selected as distinct test cases of potential distributions.

The three simulated astrophysical distributions are shown in Fig. 1, with further details given in Appendix A:

- (1) The HighSpinPrecessing population contains BHs with the most extremal spins and tilts: the majority of the population has $\chi > 0.5$ and tilts nearly in plane, corresponding to significant spin precession.
- (2) The MediumSpin population is most similar to current constraints: preferentially small to moderate spin magnitudes peaking at $\chi = 0.25$, and a wide range of tilts with a preference for alignment compared to anti-alignment.
- (3) The LowSpinAligned population has the smallest spin magnitudes, with nearly all BHs having $\chi < 0.5$. Uniquely, this population has a bimodal spin tilt angle distribution, with a larger peak at $\cos \theta = 1$ (perfect alignment) and a smaller peak at $\cos \theta = -1$ (perfect antialignment). It is therefore a test case of sensitivity to mixture models.

With these three populations, we conduct a full end-to-end injection/recovery campaign. We draw parameters describing individual GW events from each distribution, restrict to detectable events with a network optimal signal-to-noise ratio (SNR) above 10 in the LIGO Livingston, LIGO Hanford, and Virgo detectors, simulate data assuming O3 sensitivity [49], and obtain samples from the multidimensional posterior distribution of the binary parameters for each event individually. We then hierarchically model the population distribution of the simulated posteriors with parametrized population models.

The individual-event posterior sampling is conducted with the nested sampler DYNesty [50] as implemented in BILBY [51,52]. We use the IMRPhenomXPHM waveform model [27] both for simulation and recovery as it models all six spin degrees of freedom, contains higher order radiation modes, and is the least computationally expensive option available. Although more computationally expensive than approximate parameter estimation [53–55], it is essential to use full stochastic sampling for this work. As we are trying to discern subtle effects in the signals, we must properly characterize the individual-event likelihoods. Full details about parameter estimation settings are given in Appendix B. For hierarchical inference, we primarily use the Markov chain Monte Carlo sampler EMCEE [56], with some follow-up studies run with NumPyro [57,58]. The full hierarchical inference procedure is outlined in Appendix C, with the parametrized population models detailed in Appendix D.

For simplicity, the hierarchical inference ignores the azimuthal angles and in what follows use the term “spin components” to refer to the spin magnitudes and tilt angles. The parameter estimation prior and the population distribution for the azimuthal angles coincide, therefore fixing their distribution (to truth) does not incur a bias, more details are available in Appendix D.

IV. DIFFERENT SPIN MAGNITUDE AND TILT DISTRIBUTIONS CAN BE DISTINGUISHED

Using the results of the signal injection and parameter estimation campaign, we perform hierarchical inference [59–61] on events drawn from each population shown in Fig. 1 in order to reconstruct their underlying spin distributions. Population inference requires the adoption of a model for the component spin and tilt distributions. We select an analytic model in which spin magnitudes follow a

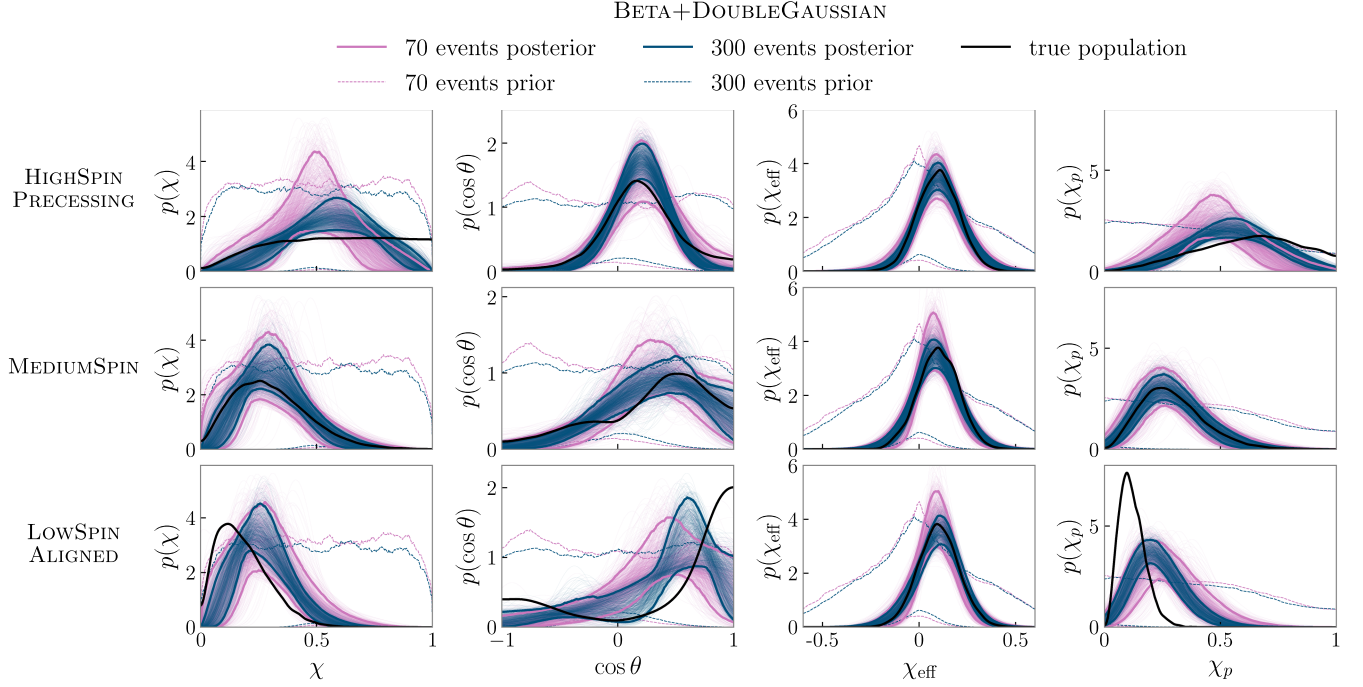


FIG. 2. Inferred distributions for various spin parameters (from left to right: spin magnitude χ , spin tilt $\cos \theta$, effective χ_{eff} spin, effective precessing spin χ_p) and the three simulated populations (top to bottom). All results are obtained with the Beta+DoubleGaussian model using 70 (pink) and 300 (navy) events. Traces correspond to draws from the population posterior and solid lines enclose 90% of the probability. The black solid line corresponds to the true, underlying population. The dashed lines show the 90% credible intervals inferred by sampling the prior on the population-level parameters, including the effective sample cut as defined in Eq. (C6).

nonsingular Beta distribution and the spin tilt angles follow a bimodal Gaussian distribution, which we hereby refer to as the Beta+DoubleGaussian; see Appendix D 2 for a full description. We choose this population model for its simplicity and similarity to common models in the literature, albeit with small modifications to target our questions of interest. While the true, underlying distributions shown in Fig. 1 were *not* explicitly drawn from the Beta+DoubleGaussian model, this model is expected to agree with each simulated population to within statistical uncertainties.¹ We do not model the spin azimuthal angle, effectively (and correctly) assuming that it is distributed according to its uniform prior.

Figure 2 shows the inferred distributions for various spin parameters for the three simulated populations under the Beta+DoubleGaussian population model. The results from a 70-event catalog are plotted in pink, and from a 300-event catalog in navy, chosen to mimic O3 and projected O4

¹We confirm that the Beta+DoubleGaussian model is a good fit to the underlying populations through a least-squares fit. The Kullback-Leibler divergences [62] between the best fit and true underlying distributions are $<10^{-4}$ for all spin magnitude distributions and <0.08 for all cosine tilt distributions. As a further check, we also run hierarchical inference on catalogs of simulated, Gaussian individual-event spin posteriors, with results discussed in Appendix G 2.

catalog size, respectively. Black, bold traces show the true underlying populations for comparison. The χ (first column) and $\cos \theta$ (second column) distributions are generated by random draws from the posteriors on the population parameters, shown in Figs. 6–8 in Appendix E.

The χ_{eff} (third column) and χ_p (fourth column) distributions are generated by (i) randomly drawing from the inferred χ , $\cos \theta$, and q distributions, (ii) calculating the effective spin parameters from these draws, and (iii) generating a Gaussian kernel density estimate. We find that *the χ_{eff} distributions are reconstructed accurately across all three populations and for both catalog sizes.*

Switching to the component spins, we can qualitatively distinguish between their distributions for each population. The spin magnitude inferred for the HighSpinPrecessing population is the widest and has the largest mean. From HighSpinPrecessing to MediumSpin to LowSpinAligned, the means and widths of the inferred distributions get progressively smaller, as is the case for the true, underlying populations. The mass ratio distributions are also successfully recovered for all three populations and two catalog sizes, as shown in Appendix E. We, therefore, *can distinguish between populations with low, moderate, and high spins when they have identical χ_{eff} distributions.*

Although we can qualitatively characterize the spin magnitude and tilt distributions among these three

populations, in some cases we cannot reliably characterize their properties accurately. Specifically, the true underlying distribution of the `LowSpinAligned` population does not lie within the 90% credible reconstructed region as measured with the `Beta+DoubleGaussian` model. The bimodality of this population's tilt angle distribution is not recovered and the inferred spin magnitude distribution has a higher mean than truth. We confirm that this mismatch between the true and inferred distributions is not *solely* driven by overly restrictive priors on the population-level parameters, plotted with dashed lines in Fig. 2. For example, the bias in the inferred spin magnitude distribution for the `HighSpinPrecessing` population occurs over a region where the prior generates a flat distribution.

Population measurements of the effective spin are more robust against bias than component spins. Even a notable mismatch between the true and recovered spin magnitude and tilt distributions for the `LowSpinAligned` population results in a precisely and accurately constrained χ_{eff} distribution. The χ_p distributions are more susceptible to inaccurate recovery in the component spins, effectively inheriting their biases. For example, for the χ_p distribution of the `LowSpinAligned` population: the inference of more in-plane spin ($\cos \theta \sim 0$) combined with the slight overestimation of the mean of the spin magnitude distribution, leads to a corresponding overestimate of the bulk of the χ_p distribution. This means that χ_{eff} , but not necessarily χ_p can be reliably characterized on a population level by component spin measurements. We additionally look at alternative definitions of χ_p (see e.g. Gerosa *et al.* [45]) and obtain over-all consistent results with the standard χ_p given in Eq. (2).

V. DIFFICULTIES OF MEASURING COMPONENT SPIN DISTRIBUTIONS

The biased reconstruction of the `LowSpinAligned` population is unexpected. The injection and recovery campaign was performed using the same waveform model, and selection effects were self-consistently handled in both signal selection and parameter estimation. Under these conditions, there is no *a priori* reason why population recovery should fail. As such, our results instead suggest a shortcoming in either the parameter estimation or population recovery stages of the analysis.

To diagnose this shortcoming, we employ a slew of checks, all of which are further elaborated upon in Appendix G. To ensure that the problem is not our hierarchical inference framework and implementation we do the following:

- (1) *Simulate Gaussian individual-event spin posteriors* (Appendix G 2): For each of the 300 events per population, we generate a series of simulated Gaussian individual-event spin posteriors with a range of

measurement errors and underlying correlations, and use these as input to hierarchical inference. In these cases, the `Beta+DoubleGaussian` population model *is* able to recover the underlying populations, as seen in Fig. 10. *Implications:* The hierarchical inference and selection effect framework is algorithmically robust, and the `Beta+DoubleGaussian` model is able to recover the true population distributions to within statistical uncertainties.

- (2) *Fix either the spin magnitude or tilt angle distribution to the truth* (Appendix G 3): When only fitting for the χ or $\cos \theta$ population and not the other, we are still unable to recover the correct distribution for the `LowSpinAligned` population; see Fig. 12. *Implications:* The observed bias in the spin magnitude and tilt angle distributions is not related to correlations between the two distributions.
- (3) *Use a different sampler for the hierarchical likelihood* (Appendix G 4): We repeat the analysis of Fig. 2 with an independently implemented hierarchical inference code that is based on NumPyro instead of EMCEE. We obtain essentially identical results, shown in Fig. 13. *Implications:* The hierarchical inference and selection effect framework is algorithmically robust.
- (4) *Fit for the mass and redshift distributions instead of fixing it to truth* (Appendix G 5): Our main results fit for the spin magnitude, spin tilt, and mass ratio distributions, while fixing the distributions of the primary mass and redshift to their true population values, given in Appendix A. Figure 13 extends these results to also fit for the mass and redshift distributions and shows the corresponding spin population posteriors, which remain unchanged. The mass and redshift distributions are recovered with no bias. *Implications:* We have not misspecified the mass or redshift distributions when fixing them to truth during hierarchical inference, nor biased results of our spin inference by neglecting to simultaneously fit for the mass and redshift distributions.
- (5) *Plot rates instead of probability distributions* (Appendix G 6): Apparent disagreement between injected and recovered probability distributions can sometimes be caused by comparing injected and recovered probability distributions, rather than differential merger rate densities. In the main text we do not infer the overall rates of black hole mergers, but only the shapes of their spin distributions. To check if neglecting the merger rate contributes to apparent disagreement between injected and recovered populations, we repeat our hierarchical inference while also fitting for the rate of black hole mergers as function of spin magnitude and tilt. This yields the results shown in Fig. 14, which remain qualitatively similar to those in Fig. 2. *Implications:* We have not

biased results of our spin measurements by failing to measure or plot absolute merger rates, rather than probability distributions.

- (6) *Exclude spin selection effects* (Appendix G 7): The selection function only negligibly affects spin magnitudes and tilt angles. To ensure that we are not incorrectly implementing the selection function in the hierarchical likelihood, we conduct hierarchical inference without including selection effects in spin. This does not impact our results, as can be seen in Fig. 15. *Implications*: The implementation of the selection function is algorithmically robust.
- (7) *Employ different methods of breaking the degeneracy in the bimodal Gaussian model* (Appendix G 7): For a bimodal distribution, some method must be imposed to break the degeneracy between the two components of the model. For the `Beta+DoubleGaussian` model, this can be done in one of three ways: imposing an ordering of the means, the widths, or limiting the mixing fraction be ≤ 0.5 . Sometimes one method of breaking the degeneracy converges better than another. As shown in Fig. 15, we find that this is not the case here and different methods perform comparably. *Implications*: Our choice of degeneracy-breaking between the two Gaussian components in our population model is not causing convergence issues.
- (8) *Run hierarchical inference on different 70-event catalog instantiations* (Appendix G 7): Finally, to get a sense of how much the specific 70 events we select from the underlying population affect hierarchical inference, we repeat the procedure with several different catalog instantiations. While there is expected variance in the results—see Fig. 16—it cannot account for the degree of mismatch seen in the bottom row of Fig. 2. Additionally, each catalog instantiation leads to a different number of per-event effective samples, which we find are not correlated to the goodness of fit. *Implications*: The observed bias does not arise from an insufficient number of per-event effective samples.

We then move towards investigating the underlying individual-event parameter estimation with the following checks:

- (1) *Sampler settings*: We run parameter estimation with a large variety of sampler settings in BILBY, and eventually adopt the standard, reviewed settings for our headline results of Fig. 2. *Implications*: Running with more aggressive sampler settings in BILBY may fix convergence problems, but this was not the case for any configurations we employed.
- (2) *Probability-probability plots* (Appendix G 1): We generate probability-probability (P-P) plots [63,64] for reweighted individual-event BILBY posteriors.

As seen in Fig. 9, the test passes. *Implications*: Either the BILBY individual-event posterior samples are unbiased, or the biases are subtle enough to not be detectable by a P-P test, as warned against in [65].

- (3) *Use a different waveform model* (Appendix G 3): We rerun individual-event inference on the same sets of events with BILBY using the `IMRPhenomXP` waveform model instead of `IMRPhenomXPHM` both for injection and recovery. Results with this waveform model are comparable or worse to that presented in the main text with `IMRPhenomXPHM`, although the bimodality of the `LowSpinAligned` population is slightly better constrained; see Fig. 11. *Implications*: The existence of bias in the measured spin magnitude and tilt distributions is not driven by our choice of waveform model, although the specific details of how that bias manifests appear to be, i.e. different waveforms yield different population-level results. This indicates that the bias may be due to individual-event sampling issues.
- (4) *Fix nonspin parameters to truth in individual-event sampling* (Appendix G 3): Finally, we conduct individual-event inference with `IMRPhenomXPHM` fixing all parameters aside from the spin magnitudes to tilt angles to truth (i.e. use delta function priors at their injected values). In this case, the `Beta+DoubleGaussian` population model is able to successfully recover the truth for all three populations, see Fig. 11. *Implications*: The added complexity going from sampling just spins to all fifteen binary parameters is a likely culprit for the biased spin magnitude and tilt angle distributions.

Although we can qualitatively tell apart the different populations in Fig. 1, our results indicate that the spin distribution of all possible BBH populations cannot necessarily be *accurately* measured under the range of analyses considered in this work. Full parameter estimation with spin precession is a technically challenging analysis. Despite conducting tens of model checking procedures, we cannot fully identify the driving source of the bias observed in Fig. 2. We hypothesize that the error is due to issues related to sampling from the high-dimensional posterior for individual events, as suggested by the final point above. If the issue with unbiased recovery is indeed due to poor convergence of parameter estimation, then it is possible that future algorithmic improvements in parameter estimation will resolve things and allow for accurate recovery.

VI. IDENTIFYING BIAS IS DIFFICULT: LIMITATIONS OF POSTERIOR PREDICTIVE CHECKS

While the `Beta+DoubleGaussian` model was able to produce qualitatively correct results for each of the three distinct

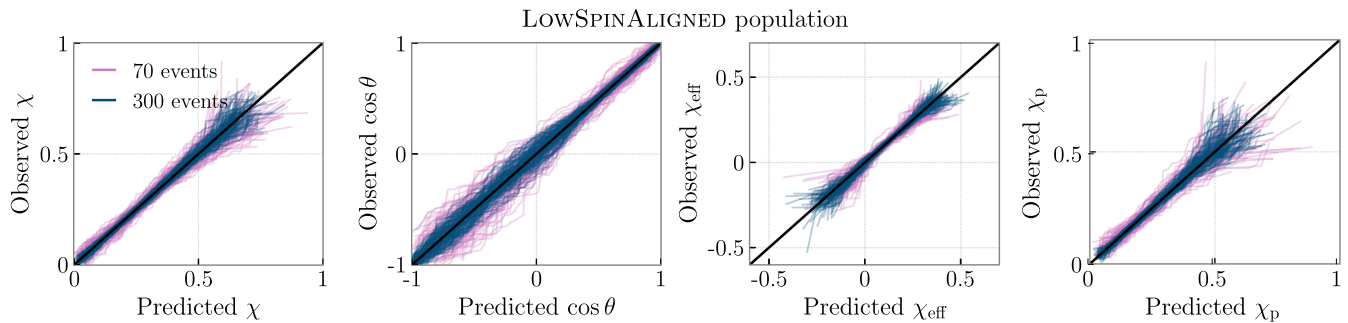


FIG. 3. PPCs for spin parameters (left to right: component spin magnitudes χ , component spin tilts $\cos\theta$, effective spin χ_{eff} , and effective precessing spin χ_p) of the `LowSpinAligned` population under the `Beta+DoubleGaussian` population model. Each trace is one catalog from the observed and the predicted populations; 100 catalogs are shown, for results with 70 (pink) and 300 (navy) events. In the absence of discrepancy between the inferred and true population distributions, the traces should on average follow the diagonal. However, here there is such a discrepancy, as can be seen in the bottom row of Fig. 2, and the traces do on average follow the diagonal, meaning that this PPC is *not necessarily* a good diagnostic tool for component spins.

populations, it was not able to successfully recover the true underlying populations. This leads us to the question: *Can commonly-used modeling diagnostics successfully identify poorly performing fits to component spin distributions?* There are multiple avenues through which a population model can fail: either the model is theoretically a good fit and for any number of reasons (e.g. those discussed in Sec. V) cannot find the truth, or the model is intrinsically a poor fit, i.e. it does not have enough flexibility to find the shape of the true, underlying distribution. Both cases induce mismatch between the true distribution and the inferred distribution, which we hope to diagnose using only the information available to us. In this section, we begin by discussing the first scenario (Fig. 3) and then the second (Fig. 4).

In reality, given the complexities of astrophysical BH spin evolution, it is almost certain that our measured distributions are in some other way discrepant with the truth; phenomenological models likely cannot perfectly reflect the underlying populations. Model checks on current data sets are then used to motivate more complicated parametric models that do not suffer from identifiable deficiencies. In parallel, nonparametric inference introduces more flexible models that are based on a large number of parameters, however those are also subject to model uncertainties and impose correlations across the population parameter space [39,41,48]. Detailed model checking remains an essential ingredient of population constraints.

For end-to-end event simulation and population recovery such as Fig. 2, we *a priori* know what the “true” underlying astrophysical distribution is. However, when dealing with real GW observations, this is, of course, not the case. We therefore diagnose the bias seen in Fig. 2 using only information available to us when dealing with real observations. To do so, we use posterior predictive checks (PPCs) that examine the predictive accuracy of

the inferred models via its ability to predict future data that are consistent with current observations. PPCs are ubiquitous in the field of GW population analyses [2,3,34,53,66].

We now look at the results from the `Beta+DoubleGaussian` model presented in Sec. IV: a case in which a population model is theoretically a good fit, but cannot find the underlying distribution accurately. A PPC for the `LowSpinAligned` population² is plotted in Fig. 3. Specifically, we plot the spin parameters predicted by the fitted model against those of the observed events. The “predicted” (horizontal axis) and “observed” (vertical axis) draws and are generated as follows:

- (1) Draw one sample from the posterior for the `Beta+DoubleGaussian` hyperparameters.
- (2) Draw one sample from the *detectable* [48,67] χ_i and $\cos\theta_i$ distribution corresponding to this hyperparameter. This is the *predicted* draw.
- (3) Draw one sample from one individual-event posterior in the catalog, reweighted to the population from Step 1, as described in Appendix F. This is the *observed* draw.
- (4) Repeat 70 times for the O3-like catalog, or 300 times for the O4-like catalog.

The predicted and observed values are sorted and plotted against each other, generating one trace in Fig. 3. We repeat this procedure 100 times to generate a collection of traces. If we have perfectly measured the true underlying distribution and in the limit of infinitely many observations, then the traces should be an exact diagonal. For a number of finite observations, the *average* of the traces should be diagonal [32,53,68–70]. As the number of observed events increases, the spread of the traces around the diagonal should decrease.

²We highlight the `LowSpinAligned` population throughout Sec. VI as it displays the largest bias under the `Beta+DoubleGaussian` model.

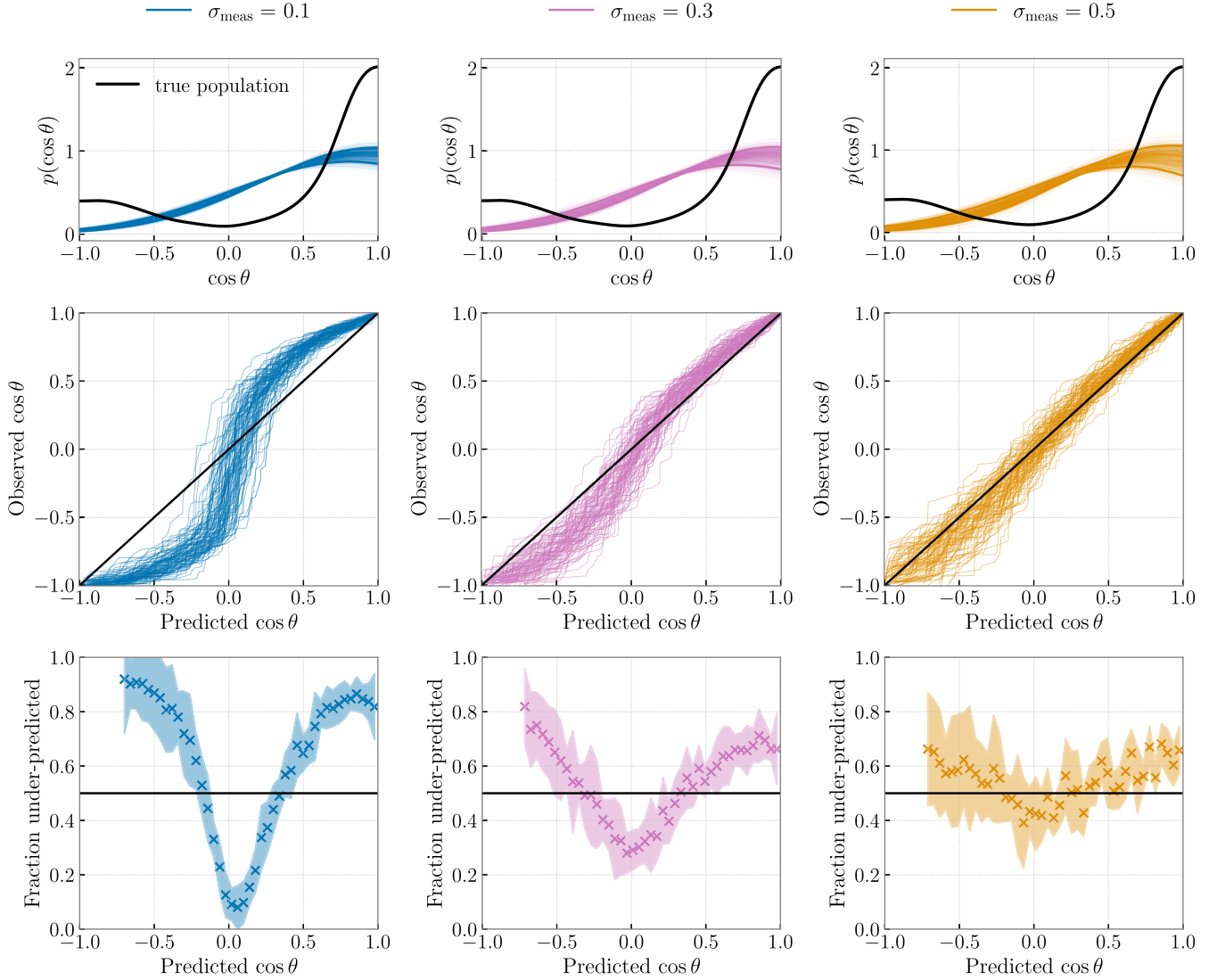


FIG. 4. Results for diagnosing model misspecification for the `LowSpinAligned` population’s $\cos\theta$ distribution under the `Beta+DoubleGaussian` model. Results from three different simulated measurement uncertainties are shown: individual-event spin measurement error of $\sigma_{\text{meas}} = 0.1$ (blue), 0.3 (pink), and 0.5 (yellow). The “true” measurement error from the `BILBY` runs averages to $\sigma_{\text{meas}} = 0.48$. All results are shown from runs done on 70-event catalogs. Top row: traces corresponding to draws from the population posterior, compared to the true population (black). Middle row: PPCs from 100 catalogs, each with 70 events. In the absence of model misspecification, the traces should on average follow the diagonal. Bottom row: Fraction of events from the posterior predictive checks with $\cos\theta$ underpredicted. The shaded regions indicate three-sigma uncertainty on each average value, marked by the crosses. In the absence of model misspecification, the error bars should encompass a horizontal line at 0.50.

For all spin parameters shown, the traces on average *do* follow the diagonal, *even though the measured population does not match the truth*. The 300 event case (navy) traces are more tightly clustered around the diagonal than the 70 event case (pink), as expected. That the traces average to the diagonal but we know the fit is poor indicates that this class of PPC, although widely used in GW population analyses, is not a sufficient diagnostic of model mismatch or inaccurate population inference in this case.

We continue to investigate the conditions under which PPCs succeed or fail by next turning to the second case discussed previously: that in which a

population model is an *intrinsically* bad fit to the underlying astrophysical distribution. We again perform population inference of simulated data, now deliberately using a model that *cannot* reproduce the injected spin distributions. In particular, we will adopt a model in which cosine tilts are described only as a single Gaussian, which we call the `Beta+DoubleGaussian` model and is given analytically in Appendix D 1. This population model *is* capable of reproducing the χ_{eff} distribution, but at the level of component spins cannot capture the bimodality present in the `MediumSpin` and `LowSpinAligned` populations.

We here wish to isolate the effect of a bad model, *without* having to worry about the shortcomings of inference per Fig. 2. As such, this time we do not perform full signal injection and parameter estimation, but instead produce mock spin magnitude and cosine tilt posteriors, allowing us to better control and understand the interplay between individual-event and population-level measurement uncertainty. We assume these mock posteriors to be Gaussian distributed with width σ_{meas} . Our procedure for generating these mock posteriors is detailed in Appendix G 2.

Results from conducting hierarchical inference using the Gaussian mock posteriors are shown in Fig. 4. The left-hand column (blue), shows results from individual-event spin posteriors with $\sigma_{\text{meas}} = 0.1$, between ~ 1 – 5 times more informative than the BILBY-produced $\cos\theta$ posteriors, which averages to $\sigma_{\text{meas}} = 0.48$. In the right-hand column (orange) are plotted results from more-realistic individual-event measurement error of $\sigma_{\text{meas}} = 0.5$; the center column (pink) is an intermediate case of $\sigma_{\text{meas}} = 0.3$.

The inferred $\cos\theta$ distributions for the LOWSPIN-ALIGNED population under the Beta+DoubleGaussian model are plotted in the top row of Fig. 4. For each of the three different individual-event measurement errors, the traces are clustered tightly, meaning that the inferred population is precisely measured, even though the model is not a good fit to the underlying population. The Beta+DoubleGaussian model is, in essence, doing its job: even with large individual event uncertainty, it identifies the mean and the overall width of the distribution very well, even though it cannot capture the full underlying bimodal structure.

We again ask the following: If we did *not* know the injected distribution, would we have been able to tell that this model is insufficient? Going further, in the case that we *can* tell a PPC fails, we are looking for an estimate of how we should amend our population model to better fit the truth. Beyond just inspecting the diagonality of PPCs by eye, we can calculate the fraction of events over/under-predicted by our model across parameter space using the slopes of the PPC traces. If the slope of a PPC trace is steeper (shallower) than the diagonal, then the model is predicting more (fewer) events in that region of parameter space than are observed. To find the slopes of each trace as a function of each parameter of interest, we perform linear regression in a small region around each point on a grid spanning that parameter. The fractions of each spin parameters underpredicted for each simulated population is then the fraction of traces with slopes shallower than the diagonal (i.e. <1). If the model is a good fit to the data, then the fraction underpredicted should be consistent with 0.5.

PPCs and the corresponding fraction of events under-predicted are shown in the middle and bottom rows of Fig. 4 respectively. Errors on the fraction underpredicted are calculated by repeating the PPC procedure ten times,

and calculating the mean (crosses) and variance (shaded region) of the results. For the $\sigma_{\text{meas}} = 0.1$ case, the PPC is inconsistent with the diagonal, meaning that here we *can* identify that inferred distribution under the Beta+DoubleGaussian model is not a good fit. The fraction of events underpredicted is correspondingly inconsistent with 0.5. For $\cos\theta \lesssim 0.25$ and $\cos\theta \gtrsim 0.75$, the fraction underpredicted is greater than 0.5, meaning that the model ubiquitously underpredicts the population in this region of parameter space. Between $0.25 \lesssim \cos\theta \lesssim 0.75$, the fraction is less than 0.5, meaning here the model over-predicts. Looking at the top left corner of Fig. 4, we can see that this is exactly the case. These results hint at how we could improve the $\cos\theta$ model: to find the truth we should allow the model to predict more events at alignment and antialignment, i.e. include a bimodality.

As the individual-event measurement uncertainty increases (left to right), the PPCs become more consistent with the diagonal, and correspondingly the fractions become consistent with 0.5. By realistic measurement uncertainty, we lose our ability to diagnose inconsistency between the underlying and measured populations using PPCs. A crucial step in generating PPCs is the reweighting of individual-event posteriors to the inferred population. If individual-event posteriors are sufficiently uninformative, then this process yields reweighted posteriors that are all essentially identical to the measured population. Thus, the “observed” and “predicted” draws will be the same, and the PPCs will be on average diagonal. *This type of PPC is therefore insufficient for weakly informative parameters as the reweighted posterior is dominated by the population rather than the individual-event likelihood.* We propose that alternative model-checking procedures must, therefore, be developed and utilized for diagnosing model bias and misspecification for poorly measured BBH parameters such as spin components.

VII. COMPARISON TO PAST WORK ON HIERARCHICAL INFERENCE WITH SPIN PRECESSION

To our knowledge, our study includes the first full, end-to-end individual-event and population-level GW injection campaign for multiple distinct populations of spin magnitudes *and* tilt angles of BBHs. Past studies performing injection-recovery campaigns for spin populations either use different waveform models and sampling implementations, and/or consider cases with reduced complexity compared to ours. Our work is in consistent with past findings, as described below.

Talbot and Thrane [71] investigated the measurability of the spin tilt angle distributions alone using an astrophysically motivated model assuming that some fraction of BBHs form in isolated binaries, while the rest form dynamically. They performed an injection and recovery campaign for spin tilts where they measured the fraction

of binary mergers with preferentially aligned versus isotropically distributed tilts, and the typical degree of spin misalignment for each BH. In their study, all simulated binaries share the same masses, distance, and spin magnitudes, chosen to be similar to LIGO’s first event GW150914 [72]. Using the waveform model IMRPhenomPv2 [73] and nested sampling implemented in LALInference [74], they found that they *are* able to constrain the parameters of the tilt-angle distributions for five different populations. Although they *do* sample over all 15 BBH parameters during individual-event inference, Talbot and Thrane [71] is most similar to the follow-up studies we present in Appendix G 3: we too are able to better recover the underlying distributions for all three of our populations when the complexity of the explored parameter space is reduced (see e.g. Fig. 12), on either an individual-event or population level.

In the context of searching for unresolved binary signals, Smith *et al.* [75] also simulated and recovered BBH spin magnitude and tilt distributions. They looked at a single population, consistent with the LIGO/Virgo O1 and O2 observations [38], and used the IMRPhenomPv2 waveform model [73] implemented in BILBY for individual-event inference. A crucial difference between this study and ours is the use of a selection function: as Smith *et al.* [75] are looking at resolved *and* unresolved binaries, they ignore selection effects entirely. Under these conditions, Smith *et al.* [75] find that the spin magnitude and tilt angle distributions are both accurately measurable. Given that their simulated populations are most similar to our MediumSpin population, this is in agreement with our findings, as we are able to well constrain the MediumSpin population for both 70 and 300 events. It is only when more complex distributions are introduced that our inference fails.

Another point of comparison between our work and others’ is on the subject of biased measurements from population model misspecification. In particular, other authors have also identified shortcomings of traditionally and widely used model checking techniques such as probability-probability plots (Appendix G 1) and posterior predictive checks (Sec. VI). Biscoveanu *et al.* [65] discussed population model bias in the mass distribution of binary neutron star populations arising from misspecification of spin priors. As part of their work, they show that a P-P check on individual-event posteriors can pass but still lead to highly biased population inference. This is in agreement with our findings.

VIII. CONCLUSIONS

In this work, we investigated the measurability of the spin magnitude and tilt angle distributions of BBH populations via GW observations. To see if realistic GW populations contain information about spin components or just the effective spin, we simulated three BBH

populations that have the same underlying effective-spin distributions, but deliberately distinct spin magnitude and tilt distributions, and on them conducted individual-event and population-level parameter estimation. We then turned to the question of whether mismatch between the injected and recovered spin magnitude and tilt distributions can be identified using only the individual-event and population-level data available to us, without knowledge of the true underlying population. Our work focuses on the three questions posed in Sec. I, the answers to which we summarize below.

- (1) *There is information in gravitational-wave signals beyond the effective-spin.* As discussed in Sec. IV, we can tell that our three different populations have different spin magnitude and tilt distributions despite their having identical χ_{eff} distributions.
- (2) *Measuring component spin distributions accurately is practically challenging.* Under standard, reviewed parameter estimation settings, we were able to accurately measure the spin magnitude and/or tilt angle distributions for some of the populations, but not all three. The bimodal tilt distribution of the LowSpinA-aligned population proved especially resistant to being accurately constrained, even when using a population model that was inherently bimodal. We employed a suite of verification methods to ensure the robustness of these results, which are enumerated in Sec. V. Notably, we find that the effective spin distribution *is*, however, accurately measured no matter the degree of mismatch in the component spin model constraints; this is not true for the effective precessing spin χ_p which remains susceptible to biased spin magnitude and tilt inference.

Although we cannot say for certain, we hypothesize that the root of the mismatch between the recovered and underlying distributions is related to a lack of convergence in individual-event posteriors, the specifics of which are subtle enough to not present themselves via a standard P-P check (Appendix G 1). We do not claim that accurately recovering component spin population distributions is *impossible* at current sensitivity, just challenging. Running BILBY with more aggressive settings, while computationally costly, may very well fix the problems presented in this work. However, unlike our injection set, real observations do not come with an answer key. If manually tuning sampler settings is a requirement to recover truth, we must be aware what these same errors could manifest in real LIGO/Virgo events.
- (3) *At current sensitivity, we cannot tell when measurements of component spin distributions are biased via the currently widely used method of posterior predictive checks.* Due to the fact that individual-event posteriors are extremely weakly informative about spin components, reweighting these posteriors to the inferred population distribution—a crucial step

in conducting posterior predictive checks—yields individual-event measurements that are all nearly identical to the inferred population itself. Nearly *any* population model can seem like a good fit to poorly constrained data, as discussed in Sec. VI.

Fishbach *et al.* [53] detailed different categories of posterior predictive checking for GW data. The most commonly used level is what we do in this work: performing consistency checks on the *true* underlying parameters of the observed data versus predicted by the model. However, one can also conduct PPCs on the *observed* parameters (e.g. max likelihood parameters) of the data versus those predicted by the model. While checks on the true parameters are susceptible to the issues related to reweighting that we discuss in Sec. III, checks on observed parameters might be more constraining. However, they are far more computationally expensive to perform, as one must generate maximum likelihood values predicted by the model: this involves either running an optimization routine or conducting mock-parameter estimation on thousands of events. While trustworthy mock-parameter estimation exists for some parameters (e.g. masses) [53–55], the imprint of spin magnitudes and tilt angles on data is more subtle and remains unincorporated into these algorithms. Developing different, more-informative methods of posterior predictive checking for poorly constrained parameters such as spin is an essential topic of future work.

Data availability. The code used to produce all results presented in this paper can be found at [76] Our individual-event and hierarchical-inference posteriors samples can be shared upon request.

ACKNOWLEDGMENTS

We thank Sylvia Biscoveanu, Jacob Golomb, Ethan Payne, and Colm Talbot for their extensive assistance with BILBY parameter estimation and other helpful comments on this work; Sophie Hourihane for essential insights about reweighting and probability-probability plots; and Salvatore Vitale for useful feedback and discussion. Furthermore, we extend thanks to Maya Fishbach, Reed Essick, Matthew Mould, Zoheyr Doctor, and Will Farr for their suggestions related to the range of verification methods performed in this paper. Finally, we thank our anonymous referees for their feedback. This work is supported by National Science Foundation Grant No. PHY-2150027 as part of the LIGO Caltech REU Program which funded Z. K. S. M. and K. C. were supported by NSF Grants No. PHY-2110111 and No. PHY-2308770. T. C. is supported by the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Futures program. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants No. PHY-0757058 and No. PHY-0823459. Software used: EMCEE [56],

BILBY (version 2.2.2) [51,52], DYNESTY (version 2.1.2) [50], NumPy [77], SciPy [78], Matplotlib [79], SEABORN [80], Astropy [81,82], JAX [83], NumPyro [57,58].

APPENDIX A: SIMULATED POPULATIONS

We simulate three populations with the same χ_{eff} but different spin magnitude χ and tilt angle θ distributions. To generate the populations, we first choose distributions of the mass ratio q and spin z component $s_{i,z} \equiv \chi_i \cos \theta_i$ to be shared in common across all three populations; this ensures the same χ_{eff} distribution. The mass ratio distribution corresponds to the median posterior value inferred with the PowerLaw+Peak model in Ref. [3], while for $s_{i,z}$ we select a Gaussian with mean 0.10 and standard deviation 0.15. This yields a Gaussian-like χ_{eff} distribution with mean 0.10 and standard deviation 0.11.

To decompose χ_{eff} into component spins, we choose a different spin magnitude χ distribution for each population and then numerically calculate the resultant $\cos \theta$ distribution implied by $p(\chi)$ and $p(s_{i,z})$. This procedure results in the three populations shown in Fig. 1. The χ distribution for the HighSpinProcessing population is uniform between $s_{i,z}$ and 1, i.e., $s_{i,z}$ values are drawn from the Gaussian described above and then a χ_i value is conditionally drawn based on each $s_{i,z}$. For the MediumSpin (LowSpinAligned) population, each χ value is drawn from a Gaussian distribution about $s_{i,z}$, truncated on $0 \leq \chi \leq 1$, with a standard deviation of 0.20 (0.05). For each population, we assume χ_1 and χ_2 are identically but independently distributed, as are $\cos \theta_1$ and $\cos \theta_2$. Finally, each spin vector’s azimuthal angle ϕ_i is drawn uniformly between 0 and 2π . Due to the different χ and $\cos \theta$ distributions, the χ_p distributions of each population differ as well.³

The astrophysical distribution of the remaining binary parameters is the same for all populations. We inject primary masses drawn from the PowerLaw+Peak model [84] with all parameters, except for m_{min} , fixed to their one-dimensional median values as found in Ref. [3]: $\alpha = 3.51$, $m_{\text{max}} = 88.21$, $\lambda_{\text{peak}} = 0.033$, $\mu_m = 33.61$, $\sigma_m = 4.72$, and $\delta_m = 4.88$ in the notation used therein. The injected mass ratio distributions for all populations are described by a power law with slope $\beta_q = 0.96$ [see Eq. (D2)], again the median inferred value from [3]. In the parametrization of the PowerLaw+Peak model, we use a population minimum mass of $m_{\text{min}} = 6M_{\odot}$ instead of $5M_{\odot}$ to set the shape of the distribution. We additionally impose a mass *cut* of $8M_{\odot}$, as restricting to higher-mass events ensures shorter analysis times. This mass cut effectively becomes the minimum mass, but we renormalize the distribution to keep the same *shape* above the cutoff mass as it

³Different component spin distributions given identical χ_{eff} and χ_p distribution can only be achieved by relaxing the assumption of identically distributed component spin magnitudes and angles.

would with $m_{\min} = 6$. Explicitly setting $m_{\min} = 8$ in the PowerLaw+Peak model would change the overall shape of the distribution to be inconsistent with the desired results in Ref. [3].

Finally, the BBH merger density rate in the source frame evolves with respect to redshift z as

$$R(z) \propto \frac{dV_c}{dz} (1+z)^{2.7}, \quad (\text{A1})$$

where V_c is the comoving volume. Distances are calculated from redshifts assuming the cosmology reported by the Planck 2013 survey [85]. All other parameters are drawn uniformly from their respective physical range. Mass and redshift distributions are plotted in Fig. 5.

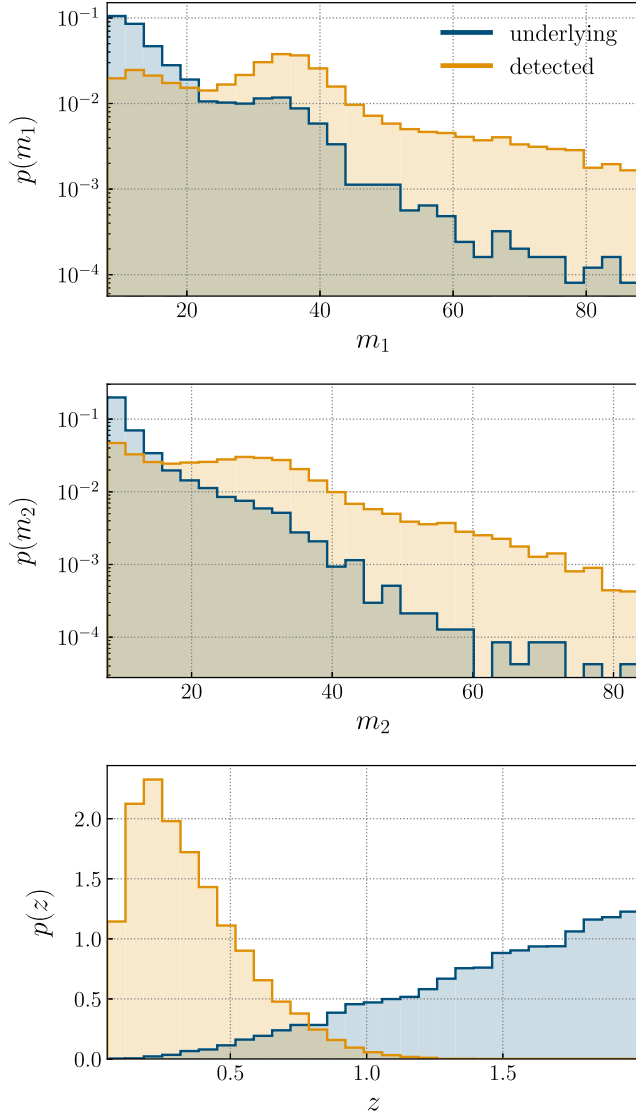


FIG. 5. The underlying (navy) and detected (orange) distributions for source-frame primary mass m_1 (top), source-frame secondary mass m_2 (middle), and redshift z (bottom) shared between all three simulated populations.

APPENDIX B: INDIVIDUAL-EVENT PARAMETER ESTIMATION

From each of the three astrophysical distributions described in Appendix A, we draw 10^5 events. We apply a network SNR [86–89] cut of 10 in the LIGO Livingston, LIGO Hanford, and Virgo detector network using the “O3 actual” power spectral densities provided in Ref. [49], and select 300 detectable events. Histograms of events from the underlying (navy) versus detectable (orange) mass and redshift distributions are shown in Fig. 5.

For each event, we simulate GW data with the IMRPhenomXPHM waveform model [27] including a Gaussian noise realization and draw samples from the 15-dimensional posterior distribution for the binary parameters using the same waveform. Specifically, we sample in detector frame component masses m_1, m_2 , spin magnitudes χ_1, χ_2 , spin tilt angles θ_1, θ_2 , the azimuthal interspin angle ϕ_{12} , the azimuthal cone precession angle ϕ_{JL} , the luminosity distance d_L , the inclination angle between the total angular momentum and the line of sight of the observer θ_{JN} , the right ascension α , declination δ , polarization angle ψ , and the time t and orbital phase φ at coalescence. We employ standard priors for all binary parameters [52], although we use a targeted chirp mass [90,91] prior of $\pm 15M_\odot$ about the injected value to reduce computational cost, which we verify does not affect results.

Simulated data assume a detector network of LIGO-Hanford, LIGO-Livingston [4], and Virgo [5], each at their O3 sensitivity [49] with a sampling rate of 2048 Hz. We analyze data in the 15–921.6 ($= 0.9 \times 2048/2$) Hz frequency range, assuming perfect knowledge of the detector calibration.

We use the nested sampler DYNESTY [50] as implemented in BILBY [51,52] under reviewed settings to stochastically sample from the individual-event posteriors. For sampler settings, we use `nlive = 1000`, `naccept = 60`, and `sample = "acceptance-walk"`. Time marginalization is turned on, while distance and phase marginalization remain off. Post-facto, we apply an optimal SNR cut of 10 on the posterior samples for consistency with our selection criteria [67].

APPENDIX C: HIERARCHICAL INFERENCE

The hierarchical inference framework used in our analysis to obtain posteriors distributions on the population parameters is implemented using the Python Markov chain Monte Carlo package EMCEE [56]. The likelihood $\mathcal{L}(\{d\}|\Lambda)$ that a catalog of N_{obs} GW events with data $\{d_i\}_{i=1}^{N_{\text{obs}}}$ arises from an underlying population π_{pop} described by parameters Λ is given by [59–61,92]

$$\mathcal{L}(\{d\}|\Lambda) \propto \prod_i^{N_{\text{obs}}} \frac{\int d\lambda \mathcal{L}(d_i|\lambda) \pi_{\text{pop}}(\lambda|\Lambda)}{\xi(\Lambda)}, \quad (\text{C1})$$

where λ_i are the parameters of the i th event in the catalog (i.e. spins, masses, etc.). In practice, we have access to the individual-event posteriors $p(\lambda_i|d_i)$ obtained with a default parameter estimation prior $\pi_{\text{pe}}(\lambda_i)$, rather than the event likelihood $\mathcal{L}(d_i|\lambda_i)$. We thus write Eq. (C1) as

$$\mathcal{L}(\{d\}|\Lambda) \propto \xi(\Lambda)^{-N_{\text{obs}}} \prod_i \int d\lambda \frac{p(\lambda|d_i)}{\pi_{\text{pe}}(\lambda_i)} \pi_{\text{pop}}(\lambda_i|\Lambda). \quad (\text{C2})$$

Additionally, rather than $p(\lambda_i|d_i)$ itself, we have a discrete set of N_i independent samples $\{\lambda_{i,j}\}_{j=1}^{N_i}$ drawn from $p(\lambda_i|d_i)$. Using the standard procedure, we approximate the integral of Eq. (C2) via a Monte Carlo average,

$$\mathcal{L}(\{d\}|\Lambda) \propto \xi(\Lambda)^{-N_{\text{obs}}} \prod_i \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{\pi_{\text{pop}}(\lambda_{i,j}|\Lambda)}{\pi_{\text{pe}}(\lambda_{i,j})}. \quad (\text{C3})$$

The detection efficiency

$$\xi(\Lambda) = \int d\lambda \pi_{\text{pop}}(\lambda|\Lambda) P_{\text{det}}(\lambda), \quad (\text{C4})$$

is the fraction of events that we would successfully detect if the population with parameters Λ is the true underlying population. Here, $P_{\text{det}}(\lambda)$ is the probability that an individual event with parameters λ is detected. As with the population likelihood, we calculate the detection efficiency with a Monte Carlo average. Given N_{inj} injected signals drawn from some reference distribution $p_{\text{inj}}(\lambda)$, the detection efficiency is

$$\xi(\Lambda) = \frac{1}{N_{\text{inj}}} \sum_{i=1}^{N_{\text{inj}}} \frac{\pi_{\text{pop}}(\lambda_i|\Lambda)}{p_{\text{inj}}(\lambda_i)}, \quad (\text{C5})$$

where the sum is over the N_{inj} injections that pass the detection criteria. We generate the set of ‘‘found’’ injections over which the Monte Carlo average is calculated in the same way that we produced catalogs of events in Appendix B. The reference $p_{\text{inj}}(\lambda)$ follows the true mass and redshift distribution (Appendix A; Fig. 5), but is uniform in spin magnitudes and isotropic in spin tilts such that we can resolve features across the full underlying spin distribution. As in Appendix B, our detection criterion is an optimal SNR greater than 10 using the waveform IMRPhenomXPHM [27] in the LIGO Livingston, LIGO Hanford, and Virgo network at O3 sensitivity [49]. We acknowledge that the optimal SNR is not a strictly accurate estimate of selection effects on real data as it is solely a function of source parameters, not detector noise. However, this approach remains formally self-consistent as long as we apply the same optimal SNR cut on posterior samples, as explained in Essick and Fishbach [67].

Following Farr [93], we account for uncertainty in the Monte Carlo integral by demanding that the effective number of independent samples

$$N_{\text{eff}}(\Lambda) \equiv \frac{[\sum_{i=1}^{N_{\text{inj}}} w_i(\Lambda)]^2}{\sum_{i=1}^{N_{\text{inj}}} [w_i(\Lambda)]^2} \geq 4N_{\text{obs}}, \quad (\text{C6})$$

where the weights w_i between the population distribution and parameter estimation prior are defined as

$$w_i(\Lambda) = \frac{\pi_{\text{pop}}(\lambda_i|\Lambda)}{p_{\text{inj}}(\lambda_i)}, \quad (\text{C7})$$

evaluated on the parameters of the *found* injections. This procedure rejects samples from regions of parameter space in which there are not sufficient injections to accurately probe. We use 200,000 injections to calculate ξ . Our results never rail against the N_{eff} cut of Eq. (C6), so we do not believe it affects our results. For further investigation, we perform a set of analyses *without* including spins when calculating N_{eff} (see Appendix G 7), under which our conclusions do not change.

In addition to including a cut on effective samples from the selection function, one can also impose a cut on the *per-event* effective samples of the posteriors used in calculating the hierarchical likelihood. Here, instead of evaluating Eq. (C7) on the found injections, it is evaluated on the *posterior samples* for every event. If any events in the catalog have an effective sample number below some threshold, then the corresponding Λ sample is tossed. In this work, we do not include any per-event N_{eff} cuts in the sampling of $\mathcal{L}(\{d\}|\Lambda)$, but calculate them post-facto as a check of Monte Carlo convergence (see e.g. Fig. 16). Other tests of Monte Carlo convergence are discussed in [94].

APPENDIX D: SPIN POPULATION MODELS

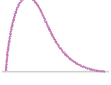
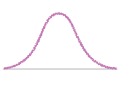
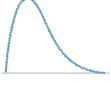
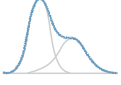
We recover the simulated populations with two models. Generically, we factorize the population models as

$$\begin{aligned} \pi_{\text{pop}}(\lambda|\Lambda) &= p(m_1|\Lambda)p(m_2|m_1, \Lambda)p(z|\Lambda) \\ &\times p(\chi_1|\Lambda)p(\chi_2|\Lambda)p(\cos \theta_1|\Lambda)p(\cos \theta_2|\Lambda), \end{aligned} \quad (\text{D1})$$

meaning, aside from the masses m_1 and m_2 , there are no correlations in the population. Our two parametrized models for the spin magnitude χ and tilt angle θ are described below. More details about the two models are provided in Table I. Spin magnitudes χ_i and tilt angles θ_i are assumed identically and independently distributed.

During hierarchical inference, we fix the distributions of primary mass m_1 and redshift z to truth, as described in Appendix A. To account for possible correlations between spins and mass ratio, although no underlying correlation

TABLE I. Details about the two component spin models we employ. Columns give the model names, an example χ and $\cos \theta$ plot, the parameters they depend on, the parameter priors, and some brief comments. The notation $U(a, b)$ means a prior uniform between a and b . For $p(\chi)$ in both models, we impose an additional prior requirement that the beta distribution shape parameters $\alpha, \beta > 1$, see Eq. (D4), making the distribution nonsingular at the boundaries. Full expressions for $p(\chi)$ and $p(\cos \theta)$ can be found in Eqs. (D3), (D6), and (D7).

Model Name	$p(\chi)$	$p(\cos \theta)$	Parameter	Prior	Comments
Beta+DoubleGaussian			μ_χ σ_χ μ_θ σ_θ	$U(0, 1)$ $U(0.07, 0.5)$ $U(-1, 1)$ $U(0.16, 0.8)$	Component spin model that cannot reproduce the simulated populations; used to study model misspecification
Beta+DoubleGaussian			μ_χ σ_χ $\mu_{\theta,1}$ $\sigma_{\theta,1}$ $\mu_{\theta,2}$ $\sigma_{\theta,2}$ f	$U(0, 1)$ $U(0.07, 0.5)$ $U(-1, 1)$ $U(0.16, 0.8)$ $U(-1, 1)$ $U(0.16, 8)$ $U(0, 5)$	Component spin model that can reproduce the simulated populations; used to study the amount of information available in component spins. See Appendix G 7 about methods of breaking the degeneracy between the two Gaussian components.

was injected, we follow Callister *et al.* [95] and simultaneously infer the distribution of binary mass ratios and spins using a secondary mass distribution of

$$p(m_2|m_1) \propto m_2^{\beta_q} (m_{\min} \leq m_2 \leq m_1), \quad (\text{D2})$$

where the power-law index β_q is a free parameter with a Gaussian prior of $\mathcal{N}(0, 3)$. The true underlying distribution has $\beta_q = 0.96$. For all other individual-event parameters, we take population distributions identical to the priors used during the original BILBY parameter estimation. Most notably for this analysis, azimuthal spins are distributed uniformly $\phi_i \in [0, 2\pi)$. All parameters aside from masses, redshift, and spin magnitudes and tilt angles are thus excluded from hierarchical inference.

1. Beta+DoubleGaussian

Following the Default model in Ref. [3], we assume that spin magnitudes χ_i are identically and independently distributed according to a Beta distribution

$$p(\chi_i|\alpha, \beta) = \frac{\chi_i^{\alpha-1} (1-\chi_i)^{\beta-1}}{B(\alpha, \beta)}, \quad (\text{D3})$$

where $B(\alpha, \beta)$ is the Beta function which ensures that the distribution is normalized to unity on $0 \leq \chi \leq 1$. Instead of sampling in the shape parameters α and β , we sample in the more familiar mean μ_χ and standard deviation σ_χ which are related to α and β by

$$\alpha = \mu_\chi \nu, \quad \beta = (1 - \mu_\chi) \nu, \quad (\text{D4})$$

where

$$\nu = \frac{\mu_\chi(1-\mu_\chi)}{\sigma_\chi^2} - 1. \quad (\text{D5})$$

We adopt uniform priors on μ_χ and σ_χ and impose an additional cut such that $\alpha, \beta \geq 1$ to keep the distribution bounded. This cut enforces $p(\chi_i|\alpha, \beta) = 0$ at $\chi_i = 0$ and 1 . Aside from the HighSpinPrecessing population at $\chi = 1$, this assumption is valid, and even there the spin model captures the overall distribution's shape.

For the tilt-angle distribution, we adopt a truncated, normalized Gaussian distribution

$$p(\cos \theta_i|\mu_\theta, \sigma_\theta) = \mathcal{N}_{[-1,1]}(\cos \theta_i|\mu_\theta, \sigma_\theta) \quad (\text{D6})$$

on the interval $-1 \leq \cos \theta \leq 1$, and fit for the mean μ_θ and standard deviation σ_θ .

2. Beta+DoubleGaussian

The Beta+DoubleGaussian model uses the same spin magnitude distribution as the Beta+DoubleGaussian, as given in Eq. (D3) and explained thereafter. The tilt angle distribution is here instead given by a mixture of two truncated normalized Gaussians

$$\begin{aligned} p(\cos \theta_i|\mu_{\theta,1}, \sigma_{\theta,1}, \mu_{\theta,2}, \sigma_{\theta,2}, f) \\ = f \mathcal{N}_{[-1,1]}(\cos \theta_i|\mu_{\theta,1}, \sigma_{\theta,1}) \\ + (1-f) \mathcal{N}_{[-1,1]}(\cos \theta_i|\mu_{\theta,2}, \sigma_{\theta,2}), \end{aligned} \quad (\text{D7})$$

to capture the multimodality of the some of the underlying distributions. We measure the means $\mu_{\theta,1}, \mu_{\theta,2}$ and standard deviations $\sigma_{\theta,1}, \sigma_{\theta,2}$ of the two Gaussians, and the mixing fraction f between them. We impose $\mu_{\theta,1} \leq \mu_{\theta,2}$ to distinguish between the two components.

APPENDIX E: DETAILED HIERARCHICAL INFERENCE RESULTS

We here show full posteriors on the hyperparameters for the Beta+DoubleGaussian component spin population model,

as given in Eqs. (D3) and (D7), for the HighSpinPrecessing (Fig. 6), MediumSpin (Fig. 7), and LowSpinAligned (Fig. 8) populations. Results for 70 (pink) and 300 (navy) event catalogs are shown in each figure. These posteriors are compared against nonlinear least-squares fit parameters

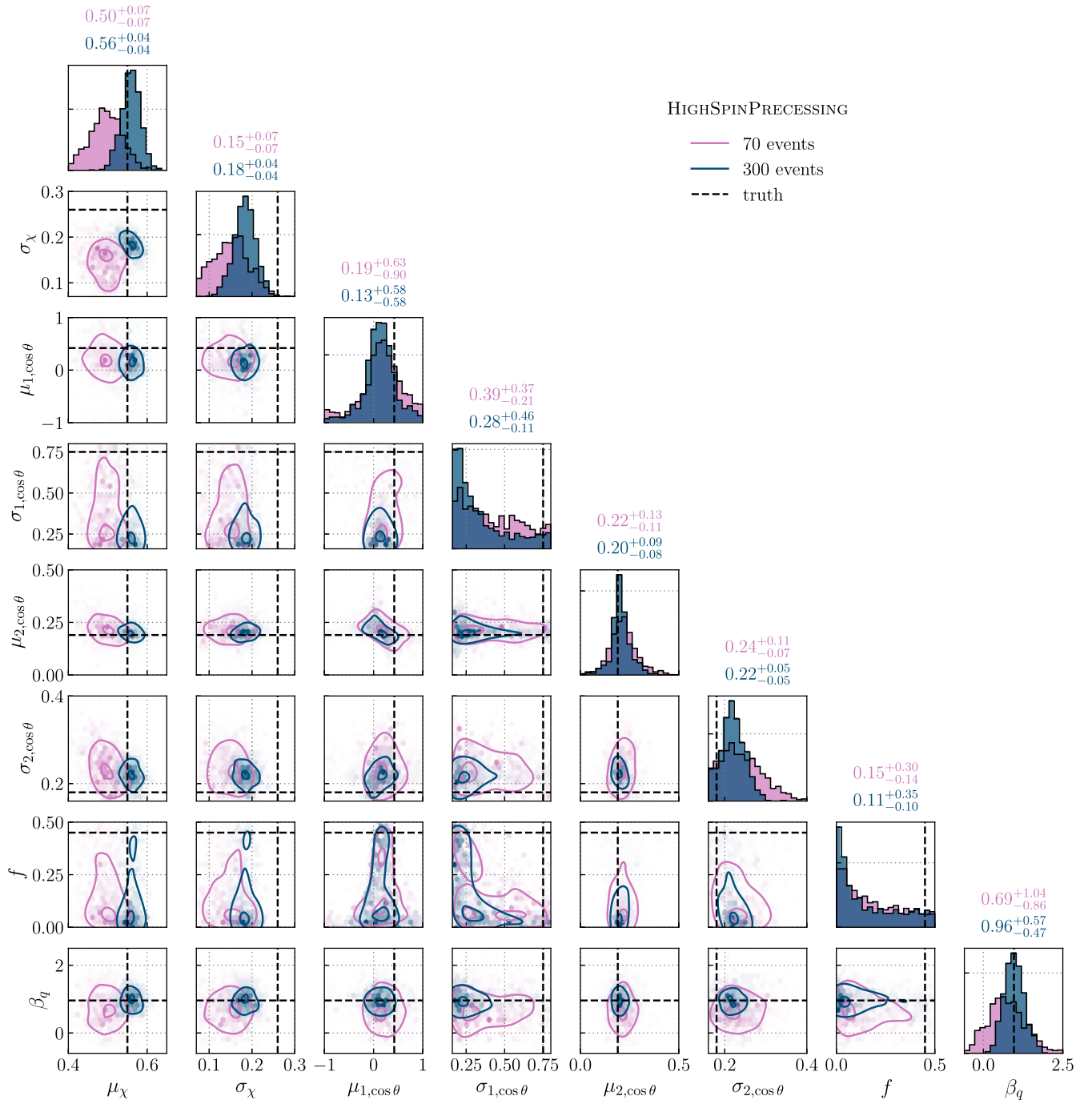


FIG. 6. The posterior distributions on the hyperparameters of the spin magnitude and tilt angle distributions under the Beta+DoubleGaussian model for the HighSpinPrecessing population for 70 (pink) and 300 (navy) event catalogs. The labels above each one-dimensional posterior give the medians and 90% credible intervals on each hyperparameter for the two different catalog sizes, while the contours in each two-dimensional posterior denote the 50% and 90% credible regions. See Table I for descriptions the hyperparameters and their priors. Black dashed lines labeled “truth” represent the theoretical best-fit parameters for the population under the Beta+DoubleGaussian model, as calculated using a least-squared fit on 50,000 draws from the population.

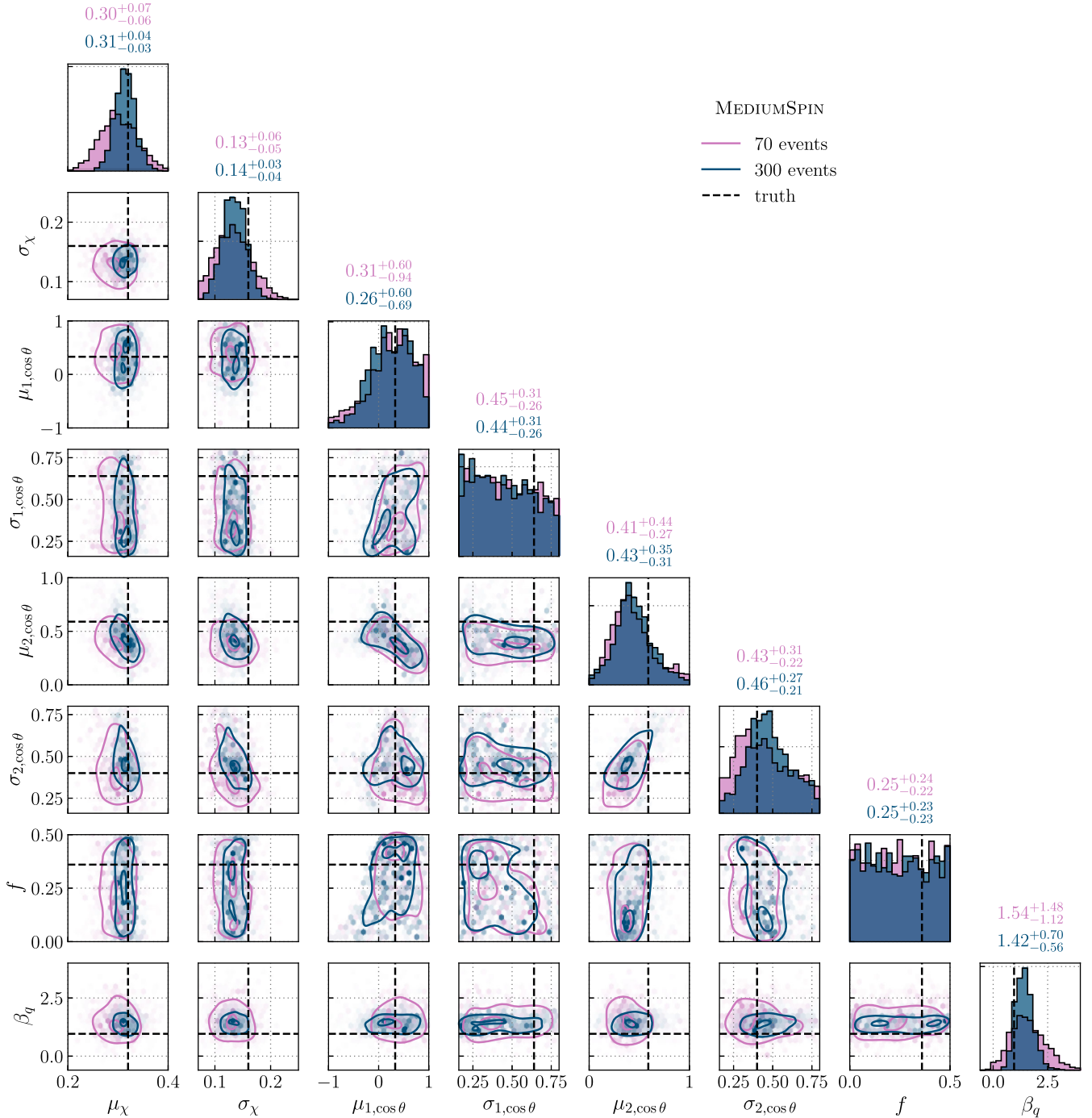


FIG. 7. Same as Fig. 6 but for the MediumSpin population. This population is recovered by the Beta+DoubleGaussian model without bias.

(black dashed) calculated from 50,000 draws per population, representing the best possible fit for the true underlying distributions within the Beta+DoubleGaussian model. Population distributions generated from draws from these posteriors are plotted in Fig. 2. For all three populations, as expected, including more events makes hyperparameter measurements more *precise*. However, adding more events does not necessarily make the results more *accurate*.

The hyperparameters of the HighSpinPrecessing population (Fig. 6) are recovered with minimal bias. While the mean μ_χ of the spin magnitude distribution of the HighSpinPrecessing population is very well constrained, its width σ_χ is slightly underestimated in the case of both the 70 and 300 event catalogs. The means $\mu_{i,\cos\theta}$ are also accurately constrained. The widths of the tilt angle distributions also seem to be underestimated, but per Fig. 2,

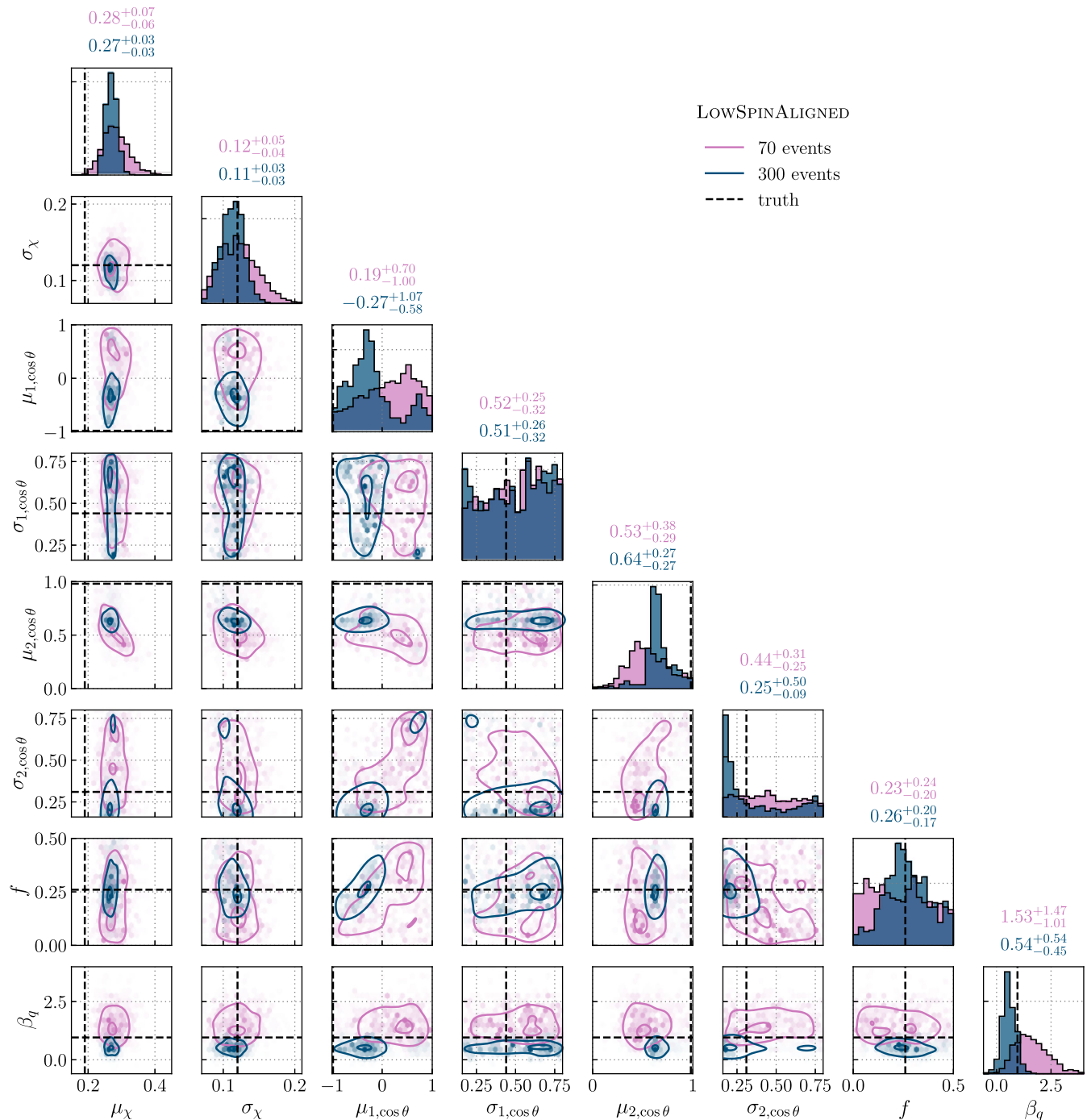


FIG. 8. Same as Figs. 6 and 7 but for the LowSpinAligned population. This population is recovered by the Beta+DoubleGaussian model with considerable bias.

the actual shape of the distribution converges on the truth. This is because—due to the allowed bimodality—different combinations of hyperparameters can lead to the same unimodal distribution.

The MediumSpin population (Fig. 7), on the other hand, is reconstructed very accurately by the Beta+DoubleGaussian model for both the 70 and 300 event catalogs. Each “true” hyperparameter either falls within

the 90% measured credible region. This is also reflected in Fig. 2—the black traces representing truth are enclosed by the 90% credible envelopes for both χ and $\cos\theta$.

Finally, the width of the spin magnitude distribution for the LowSpinAligned population (Fig. 8) is accurately constrained, but its mean μ_χ is overestimated. The most striking failure of the Beta+DoubleGaussian model is its inability to identify the bimodality of the LowSpinAligned population’s tilt

angle distribution, for either 70 or 300 events. Aside from the mixing fraction f , none of the tilt-angle distribution hyperparameters’ posteriors are consistent with truth at the 90% level.

For all three populations, the power law slope for the mass distribution is recovered within 90% credibility about the injected value of $\beta_q = 0.96$. Masses are recovered without bias by our hierarchical inference procedure; we only encountering biases when fitting for the spin populations.

APPENDIX F: REWEIGHTING INDIVIDUAL EVENT POSTERiors

Given a set of discrete sample from an individual-event posterior $p(\lambda|d_i)$ calculated with prior $\pi_{\text{pe}}(\lambda)$ and discrete samples of hyperparameters describing a population distribution $\pi_{\text{pop}}(\lambda|\Lambda)$, we can reweight the individual-event posterior to the inferred population using a two-step algorithm [32,70]. First, randomly select a hyperparameter sample from the population distribution $\Lambda_i \in \{\Lambda\}$ and calculate the following weights for each individual-event posterior sample λ_j :

$$w_j \propto \frac{\pi_{\text{pop}}(\lambda_j|\Lambda_i)}{\pi_{\text{pe}}(\lambda_j)}. \quad (\text{F1})$$

Second, select one sample $\lambda_j \in \{\lambda_j\}$ subject to the weights w_j . Repeat this process to build up a set of samples from a reweighted individual-event posterior. This procedure ensures that events are not double-counted during weighting [70].

APPENDIX G: VERIFICATION OF METHODS

To explore the origin of the bias observed in Fig. 2, we perform a number of explorations that we elaborate upon in subsequent subsections. In Appendix G 1, we generate P-P plots for reweighted individual event BILBY posteriors, which return unbiased. In Appendix G 2, we turn to hierarchical inference and explore a range of simulated Gaussian individual-event spins posteriors, rather than ones generated via stochastic sampling with BILBY. We also reduce the complexity of the parameter spaces explored on both an individual-event and population level. In Appendix G 3, we present results where we fix various combinations of parameters to their true values in both the individual-event and hierarchical inference levels. We here also discuss using a less complex waveform—IMRPhenomXP rather than IMRPhenomXPHM—which excludes higher order modes, in individual-event sampling. Appendix G 4 uses an alternate hierarchical inference code, implemented in NumPyro instead of EMCEE, and Appendix G 5 shows results from simultaneously inferring for the mass and redshift distributions along with the spins. In Appendix G 6, we look at hierarchically inferred rates

across parameter space rather than probability density functions to ensure that the normalization is not obscuring the results. Finally, other miscellaneous checks for the hierarchical inference framework and implementation are excluding selection effects in spin, trying different methods of breaking the degeneracy in the double-Gaussian tilt distribution, and looking at different 70-event catalog instantiations. Plots showing these results can be found in Appendix G 7.

1. P-P plots

A crucial assumption of hierarchical inference is that the input individual-event posteriors are themselves reliable. To test this assumption and ensure that the stochastically sampled BILBY individual-event posteriors (see Appendix B) are indeed unbiased, we perform the common diagnostic check of generating a P-P plot [63,64].

A P-P plot is generated by performing parameter estimation on events with parameters distributed according to their individual-event priors, in Gaussian noise. The percentiles, or credible intervals (CIs), at which the injections fall in their resultant one-dimensional, marginalized posterior distribution shall be uniformly distributed if parameter estimation is unbiased. In our case, where the injected distribution does *not* match the priors used in parameter estimation, reweighting (see Appendix F) to the injected distribution must be performed as a postprocessing step. Specifically, we apply an optimal SNR cut of 10 to the posteriors, and then reweight to the underlying population; this procedure is analogous to *not* applying any SNR cut and reweighting to the detected distribution.

P-P plots for spin magnitudes, spin tilt angles, masses, and redshifts for the 300 injections per simulated population are shown in Fig. 9. On the horizontal axis, the (sorted) CIs are plotted. The vertical axis shows the cumulative density function of these CIs, i.e. the frequency at which each CI occurs. This should be a diagonal line with a slope of 1 in the case of infinitely many injections: e.g. 20% of the time, the injection should fall within the lower 20% CI of its posterior. In the case of finitely many injections, these cumulative density functions should roughly fall within a $3\text{-}\sigma$ region around the diagonal, the width of which is a function of the number of events injected, as indicated by the gray lines in Fig. 9.

To check if the BILBY posteriors pass the P-P test, we look at the p values⁴ that each set of y-axis values shown in Fig. 9 is uniform. Then, we take the p values of these p values, which should also be uniformly distributed if the sampling error is random. The p values (listed in the titles of Fig. 9) for each of the three simulated populations is above the threshold of randomness expected from the seven parameters plotted ($1/7 \sim 0.143$), indicating that the BILBY posteriors pass the P-P test.

⁴We calculate p values using a Kolmogorov-Smirnov test.

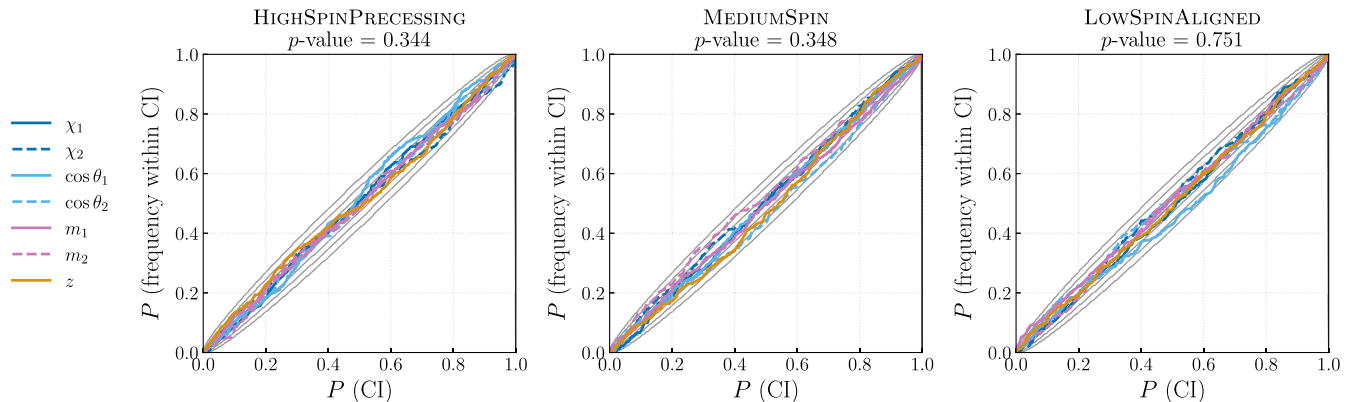


FIG. 9. P-P plots for spin magnitudes (dark blue), spin tilt angles (light blue), masses (pink), and redshifts (orange) for each simulated population (from left to right: HighSpinPrecessing, MediumSpin, LowSpinAligned). For 300 events per population, CIs (horizontal axis) are plotted against the fraction of events for which the true, injected value is recovered in that CI, stochastically sampled using the BILBY implementation of the nested sampler DYNESTY. The 1-, 2-, and 3- σ regions for 300 events are plotted in gray; all parameters stay within the 3- σ region, corresponding to the outermost gray lines. The p values for each of the three populations are greater than the threshold for seven parameters (~ 0.143), indicating that the P-P test is passed.

Though a necessary check, diagonal P-P plots are not a sufficient condition for reliable individual-event posteriors. As also seen in [65], a sampling algorithm can pass a P-P test but still result in biased hierarchical inference recovery beyond the 3- σ level. In Fig. 8, for example, the truth for the means of the spin magnitude distribution and both modes of the spin tilt distribution all lie outside of the 90% credible interval for the recovered values. It remains unclear in our case whether such discrepancies between the true and recovered populations are due to individual-event sampling issues that are not picked up by the P-P test, as was the case in [65], or further unknown biases.

2. Simulated Gaussian individual-event spin posteriors

To better understand the relation between individual-event and population-level measurement uncertainty, we generate a series of simulated individual-event spin magnitude and tilt angle posteriors for each of the same 300 GW events per population that we stochastically sample in BILBY. We take these mock posteriors to be Gaussian distributed with width σ_{meas} .

First, we generate a series of mock posteriors without any underlying spin-spin correlations with the following steps. For each of the 300 injections per population,

- (1) Take the true, injected value of each spin parameter

$$\lambda_{\text{true}} \in \{\chi_1, \chi_2, \cos \theta_1, \cos \theta_2\},$$

and from it draw an observed maximum likelihood value λ_{obs} from the Gaussian distribution $\mathcal{N}(\lambda_{\text{true}}, \sigma_{\text{meas}})$ with mean λ_{true} and width σ_{meas} .

- (2) Draw N samples from $\mathcal{N}_{[a,b]}(\lambda_{\text{obs}}, \sigma_{\text{meas}})$ where N is the number of samples in the BILBY posterior for the injection of interest and $\mathcal{N}_{[a,b]}$ is a Gaussian distribution *truncated* on $[a, b]$. For spin magnitude this

truncation is between $[0, 1]$, and for the cosine tilt angle $[-1, 1]$.

Specifically, we look at cases where $\sigma_{\text{meas}} = 0.1, 0.3, \text{ and } 0.5$, as shown in Fig. 3. In all cases, we keep the BILBY mass and redshift posteriors. Moreover, the simulated and BILBY posteriors all have the same number of samples per event.

To simulate a more realistic case, we also generate a set of mock Gaussian posteriors that do include underlying interspin correlations, with the same covariance as BILBY individual-event posteriors. For each injection, we first find the covariance of the corresponding four-dimensional BILBY posterior for $\{\chi_1, \chi_2, \cos \theta_1, \cos \theta_2\}$. We then generate a mock four-dimensional spin posterior with that same covariance using the procedure enumerated above: from truth, draw an observed maximum likelihood value; then generate a posterior by sampling a truncated Gaussian centered at that observed value. The only difference is that, instead of separately generating each one-dimensional posterior for magnitudes and tilts, we generate a four-dimensional posterior that includes correlations.

The Beta+DoubleGaussian population model *is* able to recover the underlying populations when using the simulated Gaussian posteriors, as seen in Fig. 10. This is true for the most-informative individual-event mock posteriors ($\sigma_{\text{meas}} = 0.1$; light blue dotted), the least informative ($\sigma_{\text{meas}} = 0.5$; blue dashed), and the most realistic (the posteriors with correlations, labeled “realistic σ_{meas} ”; pink solid). In all three cases, the true population lies within the 90% credible interval of the recovered region. As the measurement error decreases, the constraints get tighter around truth.

3. Reducing complexity of the explored parameter spaces

In this section, we present two simplified scenarios for individual-event sampling in BILBY, then one simplified scenario for population inference in EMCEE.

BETA+DOUBLEGAUSSIAN, 70 events with Gaussian spin posteriors

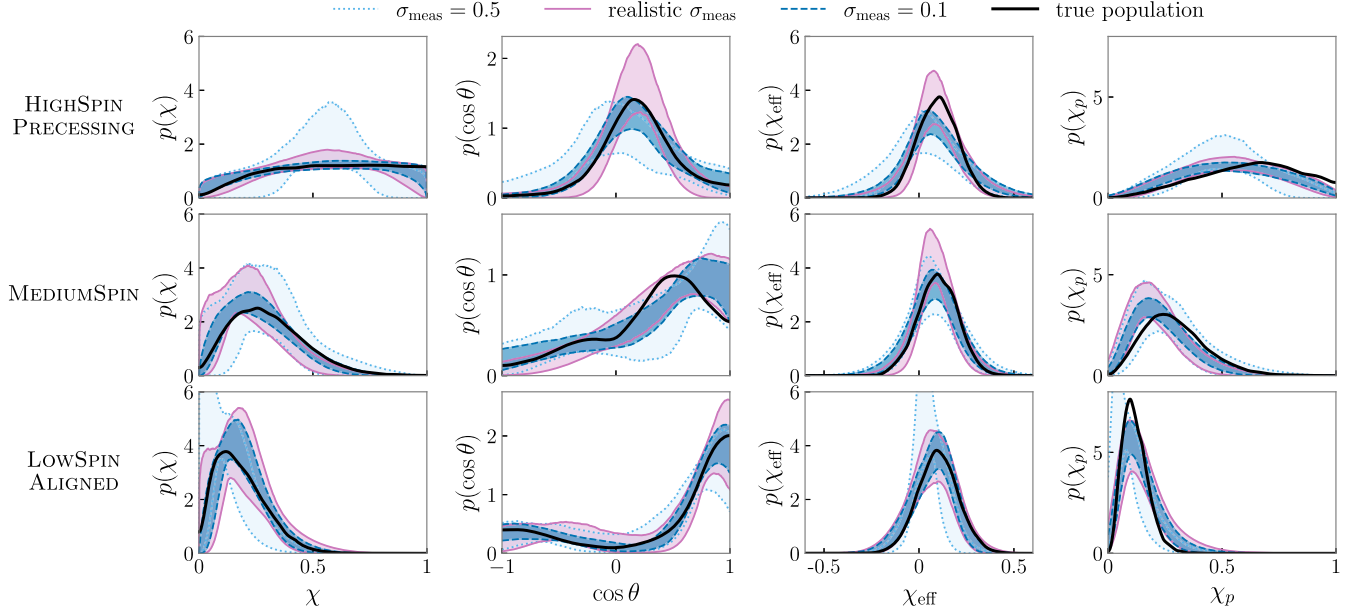


FIG. 10. Inferred distributions obtained with the Beta+DoubleGaussian model for spin magnitude χ , spin tilt $\cos\theta$, effective χ_{eff} spin, effective precessing spin χ_p , for the three simulated populations with *simulated* Gaussian individual-event spin posteriors. Results are shown from 70 event catalogs with per-event spin magnitude and tilt measurement error $\sigma_{\text{meas}} = 0.5$ (light blue dotted), $\sigma_{\text{meas}} = 0.1$ (blue dashed), and with realistic σ_{meas} and interparameter correlations taken from the BILBY posteriors (pink solid). The shaded region denotes 90% of the probability, while the black solid line corresponds to the true population.

BETA+DOUBLEGAUSSIAN with reduced dimensionality in individual-event parameter estimation

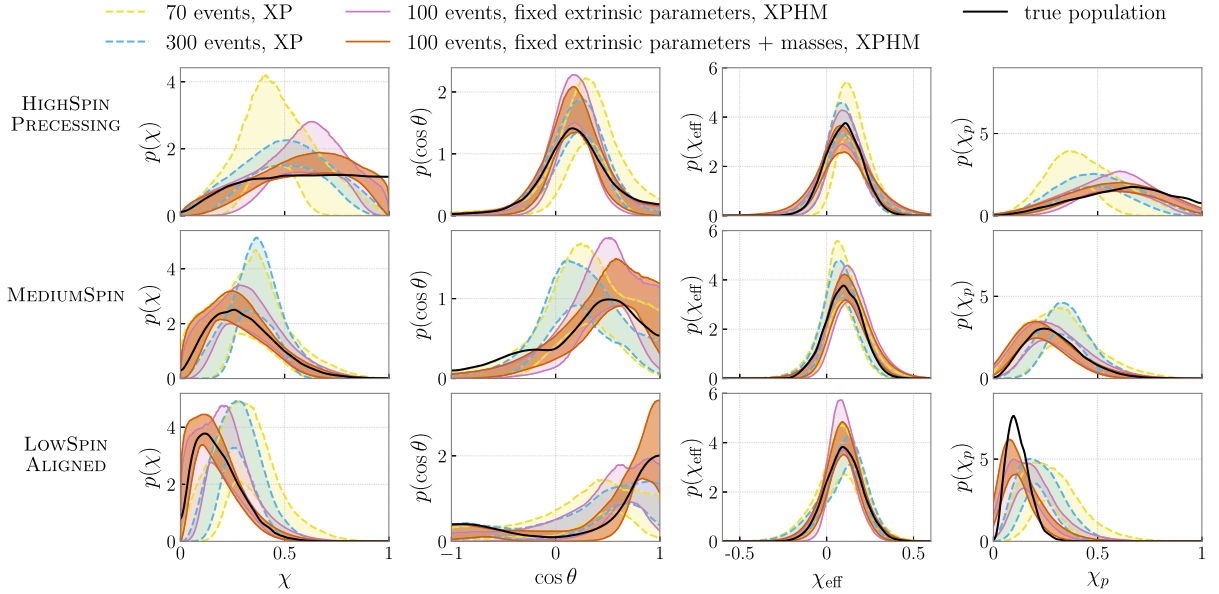


FIG. 11. Inferred distributions obtained with the Beta+DoubleGaussian model for spin magnitude χ , spin tilt $\cos\theta$, effective χ_{eff} spin, effective precessing spin χ_p , for the three simulated populations with some form of reduced complexity when the sampling of individual-event posteriors. The dashed lines show population results with the IMRPhenomXP waveform (used for both injection and recovery), which excludes higher order modes, for 70 (yellow) and 300 (blue) event catalogs. The solid lines shown populations results from 100-event catalogs using individual-event posteriors calculated with IMRPhenomXPHM, but with various parameters fixed to their true values rather than sampled over. In pink (“fixed extrinsic”), we have fixed the extrinsic parameters (i.e. everything aside from masses and spins) to their true values. In orange (“fixed extrinsic + masses”), we further restrict by *only* sampling over spin magnitudes and tilts. Of these variations, only the “fixed extrinsic + masses” individual-event posteriors yield population constraints that are improved from those shown in Fig. 2.

On the side of individual-event inference, we first reduce complexity by reducing the number of parameters sampled over. The results shown in the main text use posteriors where all fifteen dimensions of parameter-space are sampled over (see Appendix B). To simplify, we first conduct parameter estimation on all the same injections, but fixing their extrinsic parameters (i.e. everything aside from masses and spins) to the true, injected values. This yields the population constraints shown in the pink solid lines in Fig. 11 (for 100 events), labeled “fixed extrinsic parameters.” We then simplify further and additionally fix the masses and spin azimuthal angles to truth, generating the orange solid lines in Fig. 11 and labeled “fixed extrinsic parameters + masses.” Notably, the “fixed extrinsic + masses” individual-event posteriors yield population constraints that are significantly improved from those shown in Fig. 2. Since sampling convergence is more challenging as the dimensionality of the explored parameter-space increases, the trend we observe suggests that convergence might at least partially contribute to the bias.

Next, we return to sampling over all parameters (masses, spins, and extrinsic), but this time with a simpler waveform model: IMRPhenomXP [27]. Coming from the same family as IMRPhenomXPHM, the IMRPhenomXP model does not contain higher order modes, which help break degeneracies between BBH parameters. The yellow (light blue) dashed lines in Fig. 11 show the population constraints from the same 70 (300) events as Fig. 2 but with individual-event posteriors sampled with IMRPhenomXP. The recovered HighSpinPrecessing and MediumSpin populations have a worse mismatch with the truth than in Fig. 2, but the LowSpinAligned population is recovered marginally better. Higher order modes become more important to accurately constrain BBH parameters as the degree of spin precession increases. Thus, the fact that the HighSpinPrecessing population is the worst constrained by IMRPhenomXP is consistent with our understanding of the utility of higher order modes. We emphasize that these findings are unrelated to waveform systematics: we always inject and recover with the *same* waveform model.

On the population level, to reduce the complexity of the sampling, we conduct analyses where we fit for *only* the spin magnitude *or* the tilt angle distribution, while fixing the other to its true injected value. In theory, this could help identify if one or the other of these parameters was the driving factor for the mismatch between the true and recovered populations seen in Fig. 2. The inferred spin magnitude distribution for the LowSpinAligned population under the Beta+DoubleGaussian model with the tilt distribution fixed to truth is shown in blue in the top panel of Fig. 12; the bottom panel shows the inverse. These recoveries are indeed better than those in Fig. 2 (plotted in navy dashed lines for comparison), i.e. the mean of the χ distribution and mean of the larger subpopulation of the $\cos\theta$ distribution are both more accurate. However, even in this much

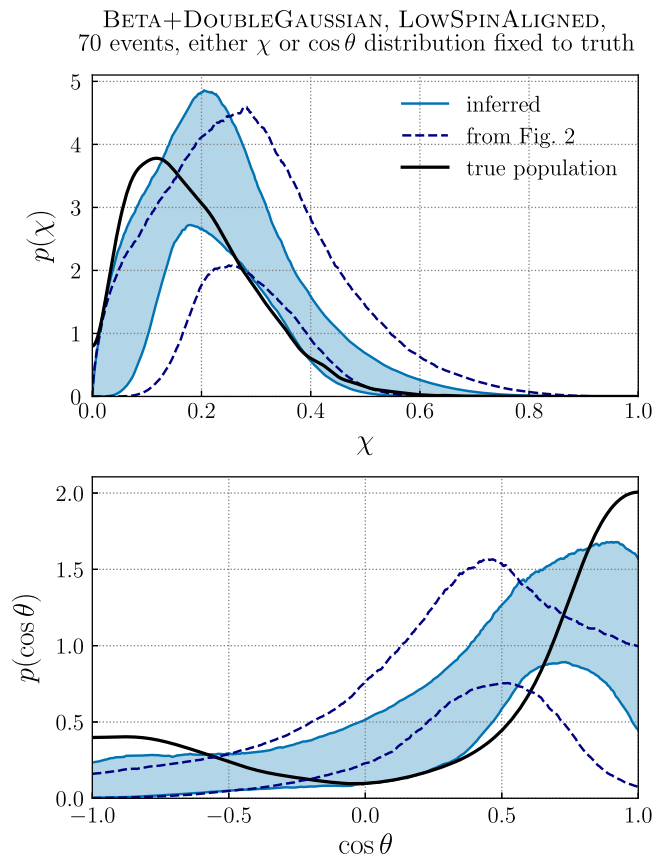


FIG. 12. Top: inferred spin magnitude χ distribution (blue shaded) for 70 events from the LowSpinAligned population under the Beta+DoubleGaussian population model with the tilt-angle distribution fixed to truth (black line in the bottom subplot). Bottom: inferred cosine of the tilt angle $\cos\theta$ distribution for the same 70 events and model with the spin magnitude distribution fixed to truth (black line in the top subplot). In both subplots, 90% credible intervals for the distributions inferred by fitting for *both* the spin magnitude and tilt distributions simultaneously, as shown in Fig. 2, are shown in navy dashed lines.

simplified version, the Beta+DoubleGaussian model again fails to recover the truth, and in particular still shows no signs of bimodality in the tilt distribution.

4. Hierarchical analysis with independent codes

It is always possible that our poor recovery of component spin distributions is simply due to an unidentified error in the code used to perform hierarchical inference. As a safeguard against this possibility, we have repeated the hierarchical analysis of the HighSpinPrecessing, MediumSpin, and LowSpinAligned populations using a second, distinct body of code. This alternate analysis code was developed entirely independently, and furthermore relies on a different stochastic sampler: whereas our main hierarchical inference results are obtained using EMCEE, this alternative performs inference using NumPyro [57,58], a probabilistic programming library implemented with JAX [83]. The NumPyro code

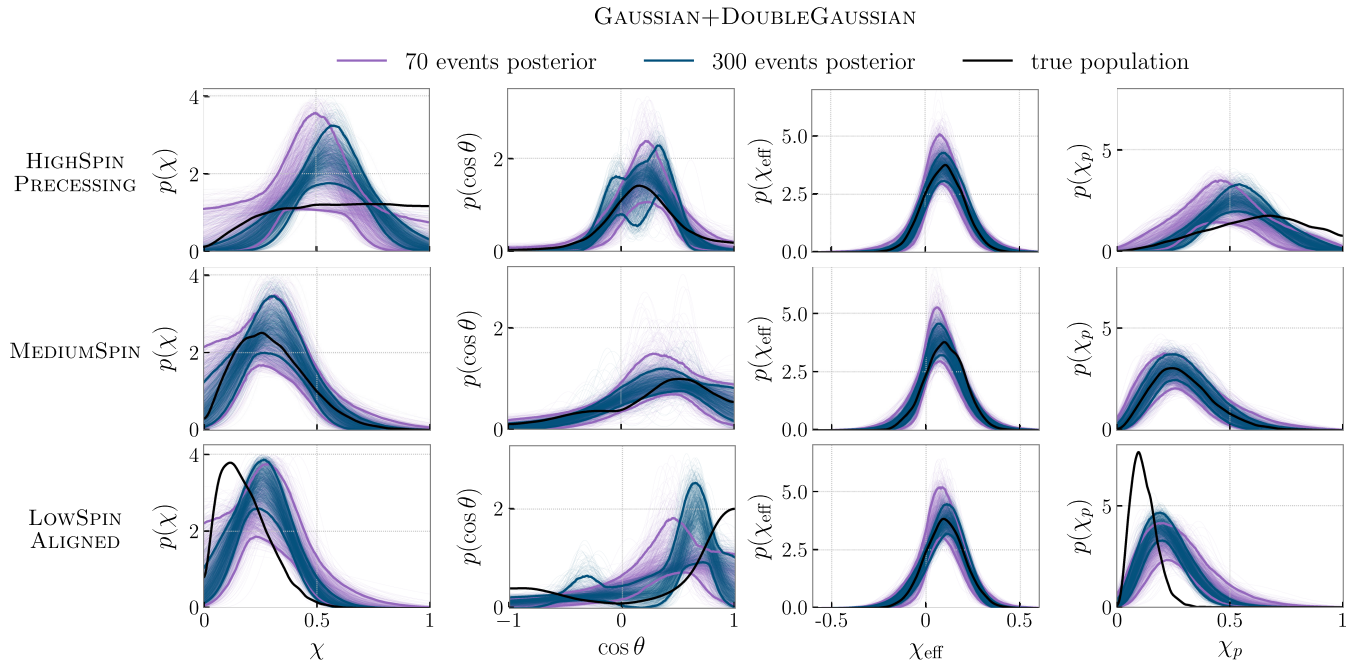


FIG. 13. As in Fig. 2, but now when additionally inferring the mass and redshift distributions of the three simulated populations in conjunction with their spin distributions. These results are furthermore produced with an entirely independent body of code, using NumPyro rather than EMCEE to stochastically sample the population likelihood. Despite these differences, the results are nearly identical to those in Fig. 2, indicating that the difficult recovery of injected component spin distributions is due neither to our choice to fix the mass and redshift distributions in the main text, nor to unidentified errors in our hierarchical inference code.

produces results nearly identical to those obtained with our EMCEE-based code. This implies that our results are not attributable to an unidentified error, unless that same error was independently introduced into two bodies of code created by two different analysts.

5. Simultaneously fitting the mass and redshift distributions

Yet another source of potential bias we investigate is the choice to fix, rather than fit, the binary black hole primary mass and redshift distributions. In principle, we do not expect significant covariance between the inferred primary mass, redshift, and component spin distributions; our simulated astrophysical populations have no underlying correlations between mass, redshift, and spin. At the same time, inferred component spins are expected to correlate strongly with the *mass ratio* distribution, which in turn can depend systematically on the choice of primary mass distribution [35,65]. Furthermore, it is known that assumptions regarding spin magnitudes can at times affect inference of the high-redshift rate of black hole mergers [2,3]. Given these possibilities, it is possible that fixing the presumed mass and redshift distributions (even fixing them to the *correct values*, as we have done) introduces bias into our spin measurements.

To check this, we repeat our inference but now hierarchically inferring the black hole mass and redshift distributions alongside the component spin distributions.

We model primary masses following the PowerLaw+Peak model [2,3] and assume that the merger rate density evolves with redshift as $(1+z)^\kappa$ for some parameter κ . This model also uses a slightly different spin magnitude model: a truncated normal distribution instead of a Beta distribution. We perform this inference using the alternative NumPyro code introduced in Appendix G 4 above. The spin distributions inferred in this case are shown in Fig. 13. The results are extremely similar to those in Fig. 2. As before, we recover the HighSpinPrecessing and MediumSpin component spin distributions reasonably well, but do not successfully measure the LowSpinAligned distributions. In this latter case, we miss (or misplace) the bimodality inherent in $\cos\theta$ and, accordingly, systematically overestimate component spin magnitudes. Once more, though, the χ_{eff} distribution is well recovered in all three cases.

6. Recovering rates and spins simultaneously

GW data are generated according to a Poisson point process, in which individual compact binaries stochastically trace an underlying *rate density* $dR/d\lambda$ of mergers across the space of binary parameters λ . When performing hierarchical inference over GW catalogs, we are formally reconstructing this rate density: measuring the “counts” of events occurring in different regions of parameter space. Often, however, we are concerned only about the shape of $dR/d\lambda$, not its normalization. In this case, it is common to

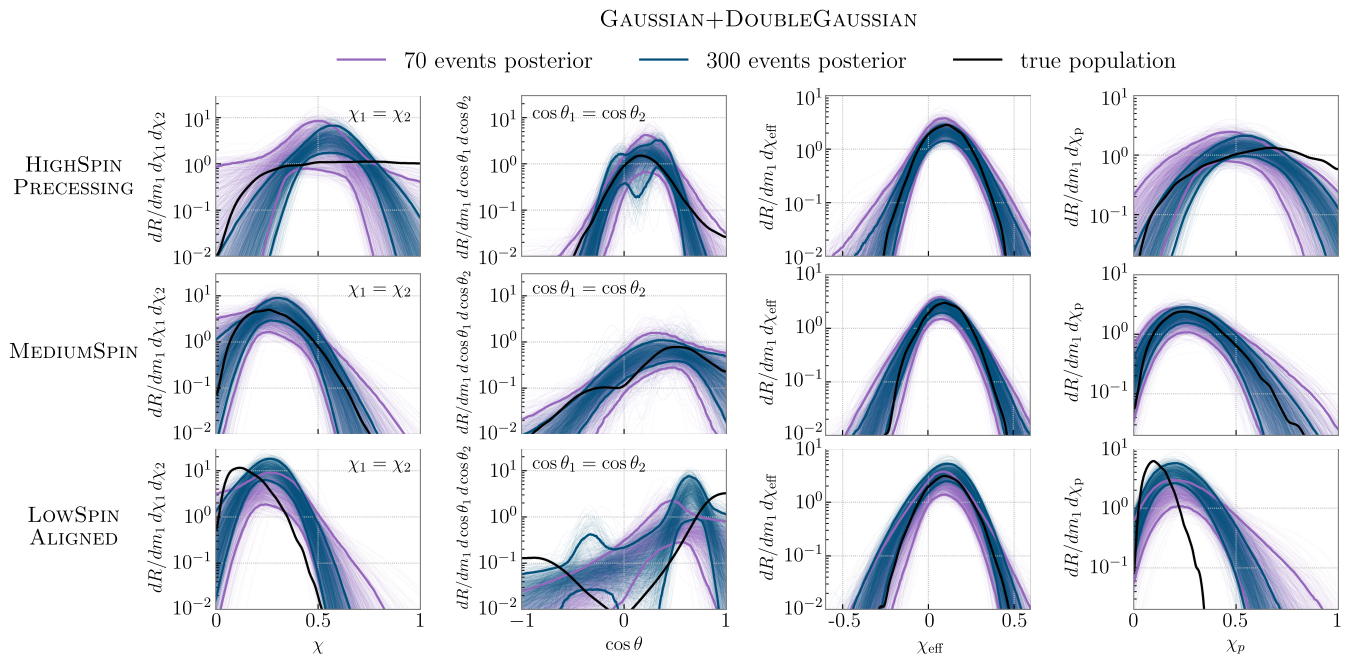


FIG. 14. As in Fig. 2, but now fitting for the absolute rate of black hole mergers as a function of spin, rather than component spin probability distributions. When producing these results, we additionally infer the black hole mass and redshift distributions, as in Fig. 13 using NumPyro rather than EMCEE. The left-most column shows the differential merger rate $dR/dm_1 d\chi_1 d\chi_2$ per unit primary and secondary spin magnitude, evaluated along the $\chi_1 = \chi_2$ line. The second column analogously shows the merger rate $dR/dm_1 d\cos\theta_1 d\cos\theta_2$ across the $\cos\theta_1 = \cos\theta_2$ line, and the last two columns give the rates and $dR/dm_1 d\chi_p$ as a function of effective spins. In all cases, rates are evaluated at $m_1 = 30M_\odot$ and $z = 0.2$. Even when plotting rates, we draw the same qualitative conclusions as originally identified in Fig. 2: the HighSpinPrecessing and MediumSpin populations are recovered well, while the rate density of the LowSpinAligned population is recovered poorly.

instead study the normalized *probability distribution* $p(\lambda)$ of source parameters. This is achieved after the fact by fitting for the merger rate but only presenting $p(\lambda)$, or from the very outset by marginalizing over and subsequently ignoring the total rate, a procedure that gives Eq. (C1).

While this procedure is usually well behaved, there do exist instances in which the choice to present the normalized $p(\lambda)$, rather than a reconstructed rate density, can yield inadvertently misleading conclusions. In some cases, models that successfully recover the correct rate density can appear to fail in recovering the correct probability density; see discussion in Sec. 5 B of Callister *et al.* [48]. Thus, when evaluating the goodness of fit of a given model, the most robust results are obtained by comparing injected and recovered *merger rates*, rather than injected and recovered probability densities.

Given this discussion, does the poor agreement between injected and recovered spin probability distributions signify a true modeling and inference failure? Or is this disagreement illusory, due to our choice to compare probability distributions rather than reconstructed merger rate densities? To check this, we repeat the hierarchical analyses of the three simulated populations but now fitting for the overall merger rate alongside the hyperparameters governing the component spin distributions.

As in Appendix G 5, we simultaneously infer the primary mass and redshift distributions. Figure 14 shows our inferred merger rates as a function of spin for each injected population. Our initial conclusions hold: when simultaneously fitting for and presenting differential merger rates, rather than probability densities, we still find that the HighSpinPrecessing and MediumSpin populations are recovered well, but we do not successfully recover the LowSpinAligned population. Hence our poor recovery of LowSpinAligned is a real effect, rather than a bias or misleading visualization related to our choice to marginalize over the absolute merger rate.

7. Other miscellaneous checks for hierarchical inference

Finally, we present results from other miscellaneous verification methods for our hierarchical inference procedure. First, we investigate different methods of breaking the degeneracy between the two Gaussian components in tilt-distribution portion of the Beta+DoubleGaussian model, see Eq. (D7). For any model defined as a mixture of multiple components, some method must be imposed to break the degeneracy these components. For a bimodal Gaussian, this can be done in three ways:

- (1) Imposing an ordering of the *means*: Assign “distribution 1” to be that with the smaller mean and “distribution 2” to be that with the larger mean.
- (2) Imposing an ordering of the *widths*: Assign “distribution 1” to be that which is narrower, and “distribution 2” to be wider.
- (3) Limiting the *mixing fraction* be ≤ 0.5 : Assign “distribution 1” to be that which contains a smaller fraction of events, and “distribution 2” to be that which contains a larger fraction.

Sometimes one method of breaking the degeneracy converges better when used in a hierarchical inference procedure than another. We find that this is not the case in this work: different methods perform identically (within sampling error), as seen in Fig. 15. Using the means (mixing fraction) of the Gaussians to break the degeneracy yields the distributions plotted in blue (orange). The results are consistent with each other. We opt to use the mixing fraction to break degeneracy throughout the bulk of this work because it is more computationally efficient.

Next, to ensure there is no misspecification in the selection function for spins, see Eq. (C5), we run hierarchical inference without any spin selection effects. Results are shown for the `LowSpinAligned` population pink in Fig. 15 for 70 (dashed) and 300 (solid) events. Selection effects in component spins are not strong, and thus are not expected to effect population inference significantly.⁵ This is indeed the case, as the distributions inferred without spin selection effects are nearly identical to those inferred with them (shown in orange). We also note that for all results shown in this work, the number of effective samples does not rail against the cut given in Eq. (C6).

Our final check is to conduct hierarchical inference on several different random 70-event catalog realizations from the 300 total events per population. Just like different Gaussian noise instantiations of the data lead to variance in individual-event posteriors, so too can random catalog instantiation lead to variance in the recovered posteriors on the *population* parameters. Some catalogs will yield a more accurately recovered population than others, just by random chance from working with finite numbers, see, e.g. Callister *et al.* [34]. Figure 16 shows some expected variance in results, but nothing corresponding to the degree of mismatch between the true and inferred tilt distributions of the `LowSpinAligned` population seen in Fig. 2. We therefore conclude that we cannot attribute bias between the injected and recovered `LowSpinAligned` population to be

⁵Selections effects are strong for masses and redshift, on the other hand. This can be seen when comparing the underlying and detected distributions in Fig. 5.

BETA+DOUBLEGAUSSIAN, 70 events, misc. tests

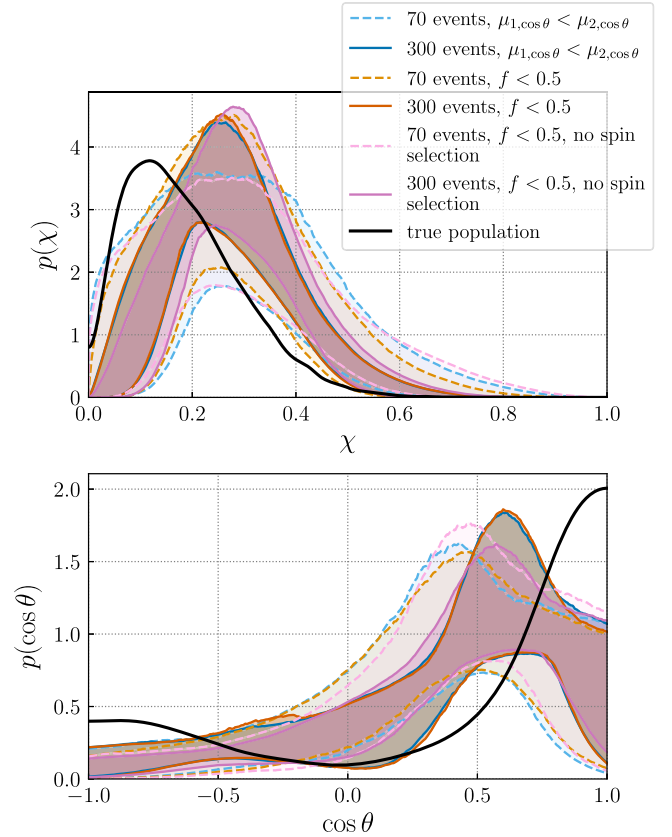


FIG. 15. Inferred distributions (blue) for spin magnitude χ (top subplot) and cosine of the tilt angle $\cos\theta$ (bottom subplot) for 70 events from the `LowSpinAligned` population obtained with the Beta+DoubleGaussian population model under various conditions. All results from 70 (300) event catalogs are plotted with dashed (solid) lines. In blue and orange we compared two methods of breaking the degeneracy between the two Gaussian subpopulations in the Beta+DoubleGaussian tilt distribution: first, restricting the mean of distribution 1 to be smaller than that of distribution 2 ($\mu_{1,\cos\theta} < \mu_{2,\cos\theta}$; blue), and second, restricting the mixing fraction to be less than one half, effectively assigning distribution 1 to be that containing less events ($f < 0.5$; orange). In orange and pink, we compare results where we do (orange) versus do not (pink) include spin selection effects (both using the $f < 0.5$ method of breaking degeneracy). None of these variations produce population constraints improved from those shown in Fig. 2; the bimodality of the tilt-angle distribution remains unrecovered.

from an “unlucky” catalog realization. Additionally, each catalog instantiation leads to a different number of per-event effective samples, see Eq. (C7) and corresponding discussion in Appendix C, which we find within a given population are *not* correlated to the by-eye goodness of fit, as seen in the rightmost column of Fig. 16. However, the `LowSpinAligned` population yields, on average, the lowest N_{eff} values and is the least accurate fit. This leads us to believe that the absence of inferred

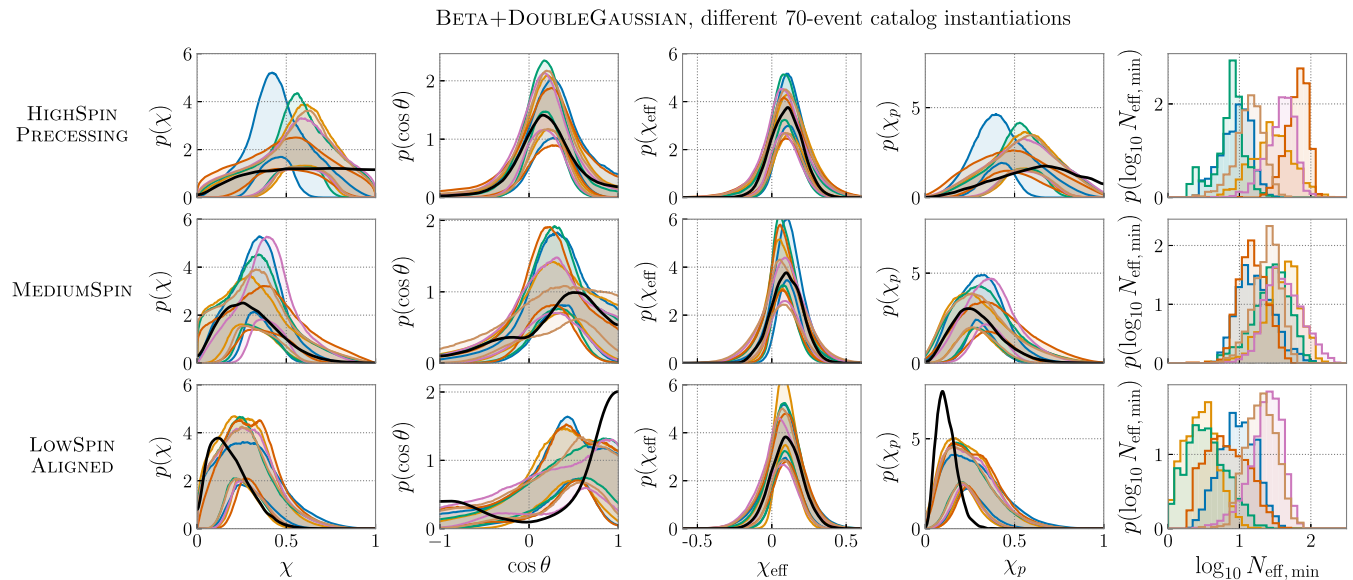


FIG. 16. Inferred distributions obtained with the Beta+DoubleGaussian model for spin magnitude χ , spin tilt $\cos\theta$, effective χ_{eff} spin, effective precessing spin χ_p , for various 70-event catalog instantiations of the three simulated populations, each plotted in a different color. The true, underlying populations are shown in black for comparison. While there is a small amount of variation, none of the catalog instantiations yields population constraints improved from those shown in Fig. 2. The rightmost column shows the minimum per-event N_{eff} over the catalog used for inference for that particular analysis [see Eq. (C7) and corresponding discussion in Appendix C]. Within a given row, the minimum N_{eff} does not correlate with the by-eye goodness of fit of the distributions in the first four columns.

bimodality is not due to the issue of our events not having enough effective samples, although the fact that the minimum event-level N_{eff} over catalog instantiations is small could be another source of imperfect recovery. As

part of future work, we plan to explore the uncertainty in difference in log-likelihood formulated in Talbot and Golomb [94] as another way to gauge whether our Monte Carlo estimations avoid bias.

-
- [1] L. Blanchet, Gravitational radiation from post-Newtonian sources and inspiralling compact binaries, *Living Rev. Relativity* **17**, 2 (2014).
 - [2] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), Population properties of compact objects from the second LIGO–Virgo gravitational-wave transient catalog, *Astrophys. J. Lett.* **913**, L7 (2021).
 - [3] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), The population of merging compact binaries inferred using gravitational waves through GWTC-3, *Phys. Rev. X* **13**, 011048 (2023).
 - [4] J. Aasi *et al.* (LIGO Scientific Collaboration), Advanced LIGO, *Classical Quantum Gravity* **32**, 074001 (2015).
 - [5] F. Acernese, M. Agathos, K. Agatsuma, D. Aisa, N. Allemandou *et al.*, Advanced Virgo: A second-generation interferometric gravitational wave detector, *Classical Quantum Gravity* **32**, 024001 (2015).
 - [6] D. Gerosa and E. Berti, Are merging black holes born from stellar collapse or previous mergers?, *Phys. Rev. D* **95**, 124046 (2017).
 - [7] C. L. Rodriguez, M. Zevin, P. Amaro-Seoane, S. Chatterjee, K. Kremer, F. A. Rasio, and C. S. Ye, Black holes: The next generation—repeated mergers in dense star clusters and their gravitational-wave properties, *Phys. Rev. D* **100**, 043027 (2019).
 - [8] S. S. Bavera, T. Fragos, Y. Qin, E. Zapartas, C. J. Neijssel, I. Mandel, A. Batta, S. M. Gaebel, C. Kimball, and S. Stevenson, The origin of spin in binary black holes: Predicting the distributions of the main observables of Advanced LIGO, *Astron. Astrophys.* **635**, A97 (2020).
 - [9] S. Vitale, R. Lynch, R. Sturani, and P. Graff, Use of gravitational waves to probe the formation channels of compact binaries, *Classical Quantum Gravity* **34**, 03LT01 (2017).
 - [10] C. L. Rodriguez, M. Zevin, C. Pankow, V. Kalogera, and F. A. Rasio, Illuminating black hole binary formation channels with spins in Advanced LIGO, *Astrophys. J. Lett.* **832**, L2 (2016).
 - [11] W. M. Farr, S. Stevenson, M. Coleman Miller, I. Mandel, B. Farr, and A. Vecchio, Distinguishing spin-aligned and

- isotropic black hole populations with gravitational waves, *Nature (London)* **548**, 426 (2017).
- [12] D. Gerosa, E. Berti, R. O’Shaughnessy, K. Belczynski, M. Kesden, D. Wysocki, and W. Gladysz, Spin orientations of merging black holes formed from the evolution of stellar binaries, *Phys. Rev. D* **98**, 084036 (2018).
- [13] I. Mandel and A. Farmer, Merging stellar-mass binary black holes, *Phys. Rep.* **955**, 1 (2022).
- [14] M. Zevin, S. S. Bavera, C. P. L. Berry, V. Kalogera, T. Fragos, P. Marchant, C. L. Rodriguez, F. Antonini, D. E. Holz, and C. Pankow, One channel to rule them all? constraining the origins of binary black holes using multiple formation pathways, *Astrophys. J.* **910**, 152 (2021).
- [15] J. Fuller and L. Ma, Most black holes are born very slowly rotating, *Astrophys. J. Lett.* **881**, L1 (2019).
- [16] Y. Qin, T. Fragos, G. Meynet, J. Andrews, M. Sørensen, and H. F. Song, The spin of the second-born black hole in coalescing binary black holes, *Astron. Astrophys.* **616**, A28 (2018).
- [17] S. S. Bavera *et al.*, The impact of mass-transfer physics on the observable properties of field binary black hole populations, *Astron. Astrophys.* **647**, A153 (2021).
- [18] N. Steinle and M. Kesden, Pathways for producing binary black holes with large misaligned spins in the isolated formation channel, *Phys. Rev. D* **103**, 063032 (2021).
- [19] M. Zevin and S. S. Bavera, Suspicious siblings: The distribution of mass and spin across component black holes in isolated binary evolution, *Astrophys. J.* **933**, 86 (2022).
- [20] M. Pürrer, M. Hannam, and F. Ohme, Can we measure individual black-hole spins from gravitational-wave observations?, *Phys. Rev. D* **93**, 084042 (2016).
- [21] S. Vitale, R. Lynch, V. Raymond, R. Sturani, J. Veitch, and P. Graff, Parameter estimation for heavy binary-black holes with networks of second-generation gravitational-wave detectors, *Phys. Rev. D* **95**, 064053 (2017).
- [22] K. Chatziioannou, G. Lovelace, M. Boyle, M. Giesler, D. A. Hemberger, R. Katebi, L. E. Kidder, H. P. Pfeiffer, M. A. Scheel, and B. Szilágyi, Measuring the properties of nearly extremal black holes with gravitational waves, *Phys. Rev. D* **98**, 044028 (2018).
- [23] S. Biscoveanu, M. Isi, V. Varma, and S. Vitale, Measuring the spins of heavy binary black holes, *Phys. Rev. D* **104**, 103018 (2021).
- [24] S. J. Miller, M. Isi, K. Chatziioannou, V. Varma, and I. Mandel, GW190521: Tracing imprints of spin-precession on the most massive black hole binary, *Phys. Rev. D* **109**, 024024 (2024).
- [25] S. Khan, F. Ohme, K. Chatziioannou, and M. Hannam, Including higher order multipoles in gravitational-wave models for precessing binary black holes, *Phys. Rev. D* **101**, 024056 (2020).
- [26] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, D. Gerosa, L. C. Stein, L. E. Kidder, and H. P. Pfeiffer, Surrogate models for precessing binary black hole simulations with unequal masses, *Phys. Rev. Res.* **1**, 033015 (2019).
- [27] G. Pratten *et al.*, Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes, *Phys. Rev. D* **103**, 104056 (2021).
- [28] S. Ossokine *et al.*, Multipolar effective-one-body waveforms for precessing binary black holes: Construction and validation, *Phys. Rev. D* **102**, 044055 (2020).
- [29] E. Racine, Analysis of spin precession in binary black hole systems including quadrupole-monopole interaction, *Phys. Rev. D* **78**, 044021 (2008).
- [30] P. Schmidt, F. Ohme, and M. Hannam, Towards models of gravitational waveforms from generic binaries II: Modelling precession effects with a single effective precession parameter, *Phys. Rev. D* **91**, 024043 (2015).
- [31] T. A. Apostolatos, C. Cutler, G. J. Sussman, and K. S. Thorne, Spin-induced orbital precession and its modulation of the gravitational waveforms from merging binaries, *Phys. Rev. D* **49**, 6274 (1994).
- [32] S. Miller, T. A. Callister, and W. Farr, The low effective spin of binary black holes and implications for individual gravitational-wave events, *Astrophys. J.* **895**, 128 (2020).
- [33] J. Roulet, H. S. Chia, S. Olsen, L. Dai, T. Venumadhav, B. Zackay, and M. Zaldarriaga, Distribution of effective spins and masses of binary black holes from the LIGO and Virgo O1–O3a observing runs, *Phys. Rev. D* **104**, 083010 (2021).
- [34] T. A. Callister, S. J. Miller, K. Chatziioannou, and W. M. Farr, No evidence that the majority of black holes in binaries have zero spin, *Astrophys. J. Lett.* **937**, L13 (2022).
- [35] K. K. Y. Ng, S. Vitale, A. Zimmerman, K. Chatziioannou, D. Gerosa, and C.-J. Haster, Gravitational-wave astrophysics with effective-spin measurements: Asymmetries and selection biases, *Phys. Rev. D* **98**, 083007 (2018).
- [36] K. Chatziioannou and W. M. Farr, Inferring the maximum and minimum mass of merging neutron stars with gravitational waves, *Phys. Rev. D* **102**, 064063 (2020).
- [37] S. Vitale, S. Biscoveanu, and C. Talbot, Spin it as you like: The (lack of a) measurement of the spin tilt distribution with LIGO-Virgo-KAGRA binary black holes, *Astron. Astrophys.* **668**, L2 (2022).
- [38] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Binary black hole population properties inferred from the first and second observing runs of Advanced LIGO and Advanced Virgo, *Astrophys. J.* **882**, L24 (2019).
- [39] J. Golomb and C. Talbot, Searching for structure in the binary black hole spin distribution, *Phys. Rev. D* **108**, 103009 (2023).
- [40] H. Tong, S. Galadage, and E. Thrane, Population properties of spinning black holes using the gravitational-wave transient catalog 3, *Phys. Rev. D* **106**, 103019 (2022).
- [41] B. Edelman, B. Farr, and Z. Doctor, Cover your basis: Comprehensive data-driven characterization of the binary black hole population, *Astrophys. J.* **946**, 16 (2023).
- [42] K. Alvi, Energy and angular momentum flow into a black hole in a binary, *Phys. Rev. D* **64**, 104020 (2001).
- [43] E. Poisson, Absorption of mass and angular momentum by a black hole: Time-domain formalisms for gravitational perturbations, and the small-hole/slow-motion approximation, *Phys. Rev. D* **70**, 084044 (2004).
- [44] D. Gerosa, M. Kesden, U. Sperhake, E. Berti, and R. O’Shaughnessy, Multi-timescale analysis of phase transitions in precessing black-hole binaries, *Phys. Rev. D* **92**, 064016 (2015).

- [45] D. Gerosa, M. Mould, D. Gangardt, P. Schmidt, G. Pratten, and L. M. Thomas, A generalized precession parameter χ_p to interpret gravitational-wave data, *Phys. Rev. D* **103**, 064067 (2021).
- [46] L. M. Thomas, P. Schmidt, and G. Pratten, New effective precession spin for modeling multimodal gravitational waveforms in the strong-field regime, *Phys. Rev. D* **103**, 083022 (2021).
- [47] P. Schmidt, M. Hannam, and S. Husa, Towards models of gravitational waveforms from generic binaries: A simple approximate mapping between precessing and non-precessing inspiral signals, *Phys. Rev. D* **86**, 104063 (2012).
- [48] T. A. Callister and W. M. Farr, A parameter-free tour of the binary black hole population, [arXiv:2302.07289](https://arxiv.org/abs/2302.07289).
- [49] LIGO Scientific and Virgo Collaborations, Noise curves used for simulations in the update of the observing scenarios paper (2020), <https://dcc.ligo.org/LIGO-T2000012/public>.
- [50] J. S. Speagle, DYNESTY: A dynamic nested sampling package for estimating Bayesian posteriors and evidences, *Mon. Not. R. Astron. Soc.* **493**, 3132 (2020).
- [51] G. Ashton, M. Hübner, P. D. Lasky, C. Talbot, K. Ackley, S. Biscoveanu, Q. Chu, A. Divakarla, P. J. Easter, B. Goncharov *et al.*, BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy, *Astrophys. J. Suppl. Ser.* **241**, 27 (2019).
- [52] I. M. Romero-Shaw *et al.*, Bayesian inference for compact binary coalescences with BILBY: Validation and application to the first LIGO–Virgo gravitational-wave transient catalogue, *Mon. Not. R. Astron. Soc.* **499**, 3295 (2020).
- [53] M. Fishbach, W. M. Farr, and D. E. Holz, The most massive binary black hole detections and the identification of population outliers, *Astrophys. J. Lett.* **891**, L31 (2020).
- [54] S. Fairhurst, C. Hoy, R. Green, C. Mills, and S. A. Usman, Simple parameter estimation using observable features of gravitational-wave signals, *Phys. Rev. D* **108**, 082006 (2023).
- [55] A. M. Farah, B. Edelman, M. Zevin, M. Fishbach, J. M. Ezquiaga, B. Farr, and D. E. Holz, Things that might go bump in the night: Assessing structure in the binary black hole mass spectrum, *Astrophys. J.* **955**, 107 (2023).
- [56] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, EMCEE: The MCMC Hammer, *Publ. Astron. Soc. Pac.* **125**, 306 (2013).
- [57] D. Phan, N. Pradhan, and M. Jankowiak, Composable effects for flexible and accelerated probabilistic programming in NumPyro, [arXiv:1912.11554](https://arxiv.org/abs/1912.11554).
- [58] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan *et al.*, PYRO: Deep universal probabilistic programming, *J. Mach. Learn. Res.* **20**, 28:1 (2019).
- [59] T. J. Loredo, Accounting for source uncertainties in analyses of astronomical survey data, *AIP Conf. Proc.* **735**, 195 (2004).
- [60] I. Mandel, W. M. Farr, and J. R. Gair, Extracting distribution parameters from multiple uncertain observations with selection biases, *Mon. Not. R. Astron. Soc.* **486**, 1086 (2019).
- [61] S. Vitale, D. Gerosa, W. M. Farr, and S. R. Taylor, Inferring the properties of a population of compact binaries in presence of selection effects, [arXiv:2007.05579](https://arxiv.org/abs/2007.05579).
- [62] S. Kullback and R. A. Leibler, On information and sufficiency, *Ann. Math. Stat.* **22**, 79 (1951).
- [63] A. G. Samantha R Cook and D. B. Rubin, Validation of software for Bayesian models using posterior quantiles, *J. Comput. Graph. Stat.* **15**, 675 (2006).
- [64] S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman, Validating Bayesian inference algorithms with simulation-based calibration, [arXiv:1804.06788](https://arxiv.org/abs/1804.06788).
- [65] S. Biscoveanu, C. Talbot, and S. Vitale, The effect of spin mismodelling on gravitational-wave measurements of the binary neutron star mass distribution, *Mon. Not. R. Astron. Soc.* **511**, 4350 (2022).
- [66] M. Fishbach and D. E. Holz, Minding the gap: GW190521 as a straddling binary, *Astrophys. J. Lett.* **904**, L26 (2020).
- [67] R. Essick and M. Fishbach, DAGnabbit! ensuring consistency between noise and detection in hierarchical Bayesian inference, *Astrophys. J.* **962**, 169 (2024).
- [68] S. Sinharay and H. S. Stern, Posterior predictive model checking in hierarchical models, *J. Stat. Plann. Inference* **111**, 209 (2003).
- [69] M. J. Bayarri and M. E. Castellanos, Bayesian checking of the second levels of hierarchical models, *Stat. Sci.* **22**, 322 (2007).
- [70] T. Callister, Reweighting single event posteriors with hyperparameter marginalization (2021).
- [71] C. Talbot and E. Thrane, Determining the population properties of spinning black holes, *Phys. Rev. D* **96**, 023012 (2017).
- [72] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GW150914: First results from the search for binary black hole coalescence with Advanced LIGO, *Phys. Rev. D* **93**, 122003 (2016).
- [73] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Simple model of complete precessing black-hole-binary gravitational waveforms, *Phys. Rev. Lett.* **113**, 151101 (2014).
- [74] J. Veitch *et al.*, Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library, *Phys. Rev. D* **91**, 042003 (2015).
- [75] R. J. E. Smith, C. Talbot, F. Hernandez Vivanco, and E. Thrane, Inferring the population properties of binary black holes from unresolved gravitational waves, *Mon. Not. R. Astron. Soc.* **496**, 3281 (2020).
- [76] <https://github.com/simonajmiller/measuring-bbh-component-spin>.
- [77] C. R. Harris *et al.*, Array programming with NumPy, *Nature (London)* **585**, 357 (2020).
- [78] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in Python, *Nat. Methods* **17**, 261 (2020).
- [79] J. D. Hunter, Mplotlib: A 2d graphics environment, *Comput. Sci. Eng.* **9**, 90 (2007).
- [80] M. L. Waskom, SEABORN: Statistical data visualization, *J. Open Source Software* **6**, 3021 (2021).
- [81] Astropy Collaboration, Astropy: A community Python package for astronomy, *Astron. Astrophys.* **558**, A33 (2013).
- [82] Astropy Collaboration, The Astropy Project: Building an open-science project and status of the v2.0 Core Package, *Astron. J.* **156**, 123 (2018).

- [83] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary *et al.*, JAX: Composable transformations of Python+NumPy programs (2018).
- [84] C. Talbot and E. Thrane, Measuring the binary black hole mass spectrum with an astrophysically motivated parameterization, *Astrophys. J.* **856**, 173 (2018).
- [85] P. A. R. Ade *et al.* (Planck Collaboration), Planck 2013 results. XVI. Cosmological parameters, *Astron. Astrophys.* **571**, A16 (2014).
- [86] K. S. Thorne, Gravitational radiation, in *Three Hundred Years of Gravitation* (Cambridge University Press, Cambridge, 1987), pp. 330–458.
- [87] L. S. Finn and D. F. Chernoff, Observing binary inspiral in gravitational radiation: One interferometer, *Phys. Rev. D* **47**, 2198 (1993).
- [88] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, FINDCHIRP: An algorithm for detection of gravitational waves from inspiraling compact binaries, *Phys. Rev. D* **85**, 122006 (2012).
- [89] R. Essick, Semianalytic sensitivity estimates for catalogs of gravitational-wave transients, *Phys. Rev. D* **108**, 043011 (2023).
- [90] P. C. Peters, Gravitational radiation and the motion of two point masses, *Phys. Rev.* **136**, B1224 (1964).
- [91] L. Blanchet, T. Damour, B. R. Iyer, C. M. Will, and A. G. Wiseman, Gravitational-radiation damping of compact binary systems to second post-Newtonian order, *Phys. Rev. Lett.* **74**, 3515 (1995).
- [92] E. Thrane and C. Talbot, An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models, *Pub. Astron. Soc. Aust.* **36**, e010 (2019).
- [93] W. M. Farr, Accuracy requirements for empirically-measured selection functions, *Res. Notes Am. Astron. Soc.* **3**, 66 (2019).
- [94] C. Talbot and J. Golomb, Growing pains: Understanding the impact of likelihood uncertainty on hierarchical Bayesian inference for gravitational-wave astronomy, *Mon. Not. R. Astron. Soc.* **526**, 3495 (2023).
- [95] T. A. Callister, C.-J. Haster, K. K. Y. Ng, S. Vitale, and W. M. Farr, Who ordered that? unequal-mass binary black hole mergers have larger effective spins, *Astrophys. J. Lett.* **922**, L5 (2021).