SPINS

INT-PUB-25-016

# The Long Road to Alignment: Measuring Black Hole Spin Orientation with Expanding Gravitational-Wave Datasets

Salvatore Vitale* and Matthew Mould

*LIGO Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

*Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA and Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

(Society of Physicists Interested in Non-aligned Spins, SPINS)†

(Dated: June 12, 2025)

Measuring the distribution of spin tilts—the angles between the spin vectors and the binary orbital angular momentum—in stellar-mass binary black holes detected by LIGO-Virgo-KAGRA would provide valuable insight into their astrophysical origins. Analyses of the 69 binary black holes detected through LIGO-Virgo-KAGRA's third observing run yielded model-dependent conclusions, particularly regarding whether the spin tilt distribution exhibits a peak near alignment, as expected for binaries formed in galactic fields. In this work, we simulate populations of up to 1500 binary black hole systems with parameters consistent with the default GWTC-3 analysis, while introducing a correlation that favors small spin tilts for binaries with mass ratios near unity. We find that: (a) spurious peaks away from perfect alignment are possible even with catalogs of up to 300 sources; (b) establishing a definitive peak at alignment remains difficult even with 1500 detections; (c) integrated measurements – such as the fraction of events with tilt angles smaller than $10^\circ$ or greater than $90^\circ$ – are more robust and should be preferred, achieving relative 90% credible uncertainties of $\sim 20\% - 80\%$ with 1500 sources; and (d) even with the largest simulated catalogs, evidence for a mass ratio–tilt correlation remains inconclusive. Our results suggest that identifying the formation channels of merging black holes using spin tilts will remain challenging, but that model-independent measurements may yield more informative insights over model parameters themselves.



## I. INTRODUCTION

The field of gravitational-wave (GW) astrophysics will become ten years old in September 2025. After the momentous discovery of GW150914 [1], the LIGO [2]-Virgo [3]-KAGRA [4] (LVK) collaboration and others have published on the detection of nearly 100 binary black hole (BBH) binaries [5–7]. With another $\sim$ 200 significant compact binary mergers [8] - most of which are BBHs - reported in low-latency in GCNs during the fourth observing run (O4) and yet to be published [9], population analyses are likely to become more and more constraining. These studies aim at characterizing the properties of the underlying astrophysical population - or populations - of the stellar-mass BBHs being detected in LVK data [10, 11]. The ultimate goal would be to understand how many such populations exist across the universe, their merger rates over redshift, and the distribution of BBH parameters (masses, spins, eccentricity, though eccentricity is much harder to measure [12]) arising from each population. In practice, one deals with the inverse problem: given a set of $N \gg 1$ detected BBHs, for each of which a noisy measurement of masses, spins, etc. is obtained, what can be said about the astrophysical populations that produced them?

To attempt answering this question, predictions about the distributions of some or all parameters from each population (also known as formation channel) should be available. Those can be either partial reasonable expectations on only one or a subset of the parameters or very detailed distributions of all parameters and their correlations. These latter are generated by so called population synthesis suites: end-to-end simulations that may yield the distribution of all intrinsic parameters and their correlations [13–18]. The analyst must thus decide what level of expectations to fold in when setting up their models for the astrophysical distribution of BBH parameters.

The most conservative approach is to use flexible models [19–32]. This approach has the benefit of minimizing the risk of biases but comes at the cost of larger statistical uncertainties driven by the large number of model's parameters. It also usually yields constraints about the properties of the *overall* set of sources, without characterizing eventual subpopulations. The next most conservative approach is to use parametric families of functions to model some or all of the source properties (e.g., Refs. [33–40]). This is the approach that was historically followed first, as it is simple to implement and has the advantage that at least some of the model's parameters can approximate astrophysical quantities of interest. More recently there have been proposals to use directly the output of populations synthesis codes as the model that enters the population analysis [41–45]. The prospects of using population synthesis results as models for the analysis of GW data is intriguing, as they could constrain the very parameters that affect binary evolution (e.g. the properties of the progenitor stars initial mass function). At the same time, limitations exist in the complexity and availability of such population synthesis simulations everywhere in the relevant parameter space. Using limited informa-

* salvo@mit.edu

† sites.mit.edu/spins

tion can naturally lead to biases [46, 47]. In practice, one might use a combination of all the methods above in various steps of the analysis, or for different subset of parameters [45, 48].

Among the tracers of BBHs formation channels there are BH spins [33, 35, 49–57]. While predictions on the expected spin magnitude have changed over the years and are hard to make in a solid fashion (see for example the introduction of Ref. [58] and references therein), a rather simple prediction can be made about their direction. That is that BBHs formed dynamically should have spins that are randomly oriented, since no direction is preferred, whereas black holes formed in galactic field stellar binaries should have spins preferentially aligned with the binary orbital angular momentum [59, 60]. The exact degree of alignment is not known, as it depends on the poorly-understood details of the supernovae explosions that generate the two BHs, for example, the geometry and amount of the fallback material, and on the orbital separation at the time of the supernovae [61–63].

This expectation can be folded in a simple model for the spin angles (called tilts and indicated with the letter $\tau$ in this paper), that includes a mixture of two components: one isotropic component and a preferentially-aligned component. In Ref. [33] this preferentially-aligned component was a Gaussian in cosine tilt space centered at $\cos\tau = 1$ and with a fixed width. Ref. [35] later extended the model by making the width of the Gaussian a parameter measured from the data. This is the model that was used in the default spin analyses of the LVK up to their catalog GWTC-3 based on the end of the third observing run (O3b). In their GWTC-3 paper [40], the LVK finds that the data are consistent with the tilt distribution having a peak at aligned spins but also with a broad distribution (they find that $44^{+6}_{-11}\%$ of BHs have $\cos\tau \leq 0$[1]). Ref [64] extended the spin tilt model by allowing the location of the preferentially-aligned component to be measured from the data and re-analyzed GWTC-3 with it. They found that while the data is not inconsistent with a peak at $\cos\tau = 1$ it does not require it either. In fact, with this extended model they obtain a mild preference for a broad peak in the $\cos\tau$ distribution *away* from unity. This unexpected result was corroborated in that same paper using different parametric models [64] and by other independent analyses [21, 24–26, 65].

The challenges of measuring the magnitude and tilt of individual BHs with GWs has been known for a long time [66–70], and naturally these uncertain measurements on an event-by-event basis propagate at the population level, making the choices of models, priors and analyses details more important [64, 71]. Indeed, Ref. [72] investigated the extent to which GW observations constrain the full spin distributions of BBHs beyond the commonly used effective spin parameters. Using simulated populations with identical effective spin distributions but differing component spin magnitudes and tilt angles, they found that while gravitational waves do encode information about full spin vectors, this information is extremely difficult to extract in practice.

In this work we focus on the measurability of the astrophysical spin tilt distribution using simulated BBH populations. Our "true" population is chosen to be consistent with what measured in GWTC-3 but we also endow the population with a correlation between BBHs mass ratio and spins (which does leave it consistent with GWTC-3). We consider catalog sizes of up to 1500 sources. That is, we attempt to make predictions about the evolution of this measurements for the next several years. These sources are analyzed with several of the models proposed by Ref. [64], in order to assess model-dependence and enable model selection. Our main findings are that:

- Even if the true $\cos\tau$ distribution peaks at 1, spurious peaks away from unity are not impossible with catalogs including as many as 300 sources. However, they become increasingly less likely as the catalog size increases. Should a peak away from $\cos\tau = 1$ still be observed in O4 data (which could include $\simeq 300$ BBHs, based on public alerts [8]) it might still go away as more sources are added.

- Even with 1500 BBH sources, some of the measurements of key parameters — such as the fraction of events in the non-isotropic component — are very uncertain, and may depend on the exact model being used, especially for our most elastic model.

- While it is tempting to focus on marginalized posteriors of the model parameters, as some of those can be directly connected to interesting astrophysical quantities, it is better to work directly with the posterior predictive distribution (PPD) of the relevant source parameter or with integrated quantities obtained from the PPD. Given the complexity of these models, looking at marginalized posteriors of individual parameters can be challenging and might belie the whole story.

- For our simulated population, the true fraction of sources with $\cos\tau \leq 0$ is 38.9% and the true fraction of sources with tilts within 10° from perfect alignment is 1.0%. Using a mixture model with an isotropic spin component and a Gaussian component with both location and width measured from the data, we measure these two fractions to be $37.1^{+7.4}_{-7.8}\%$, $41.8^{+3.9}_{-4.6}\%$, $40.7^{+2.8}_{-2.2}\%$ and $1.0^{+0.3}_{-0.3}\%$, $0.9^{+0.2}_{-0.2}\%$, $0.9^{+0.1}_{-0.1}\%$ with catalogs including $N = 150$, 500 and 1500 sources, respectively.

- Irrespective of the catalog size, we cannot reveal in a definitive way the existence of the mass ratio–spin tilt correlation. Both posterior-based analyses and evidence-based analyses are inconclusive. In

[1] As shown in our Fig 11 and Tab. IV, the prior for this fraction peaks in the same region.

fact, the simplest model–that does not allow for the existence of this correlation–is slightly preferred for all catalog sizes, with most other models achieving a similar evidence.

The rest of the paper is structured as follows: in Sec II we discuss in detail the parameters of the simulated BBH populations; in Sec. III we review the technical details of hierarchical Bayesian analysis; in Sec. IV we list the models used in the paper. Results are reported in Sec. V. Sec. VI discusses our findings and future prospects.

## II. SIMULATED POPULATION

We simulate BBH mergers from a population that is consistent with GWTC-3 data [40]. Specifically, the joint distribution for the source-frame primary mass ($m_1$) and mass ratio ($q$) is the `Power Law + Peak` model [34] with hyper parameters: $\alpha_{m_1} = 3.4, m_{\rm min} = 5M_\odot, m_{\rm max} = 87M_\odot, m_{\rm lam} = 0.04, m_\mu = 34M_\odot, m_\sigma = 3.6M_\odot, \delta m = 4.8M_\odot, \beta_q = 1.1$. The spin magnitudes are assumed to be independent and identically distributed from a beta distribution [73], with parameters $\alpha_\chi = 1.67, \beta_\chi = 4.43$. The redshifts are draws from the `power law` redshift model [36] with slope $\lambda_z = 2.73$. Fig. 17 in App. A shows the hyper posteriors for the default LVK model [40] for GWTC-3 (as run by Ref. [64]) in blue, and mark the values of the hyper parameters we used to simulate our sources with yellow lines.

The cosine of the black hole spin tilt angles are drawn from a distribution that is a mixture of an isotropic distribution, and a Gaussian peaking at $\cos\tau = 1$ (i.e., spin vector aligned with the angular momentum). In order to verify if and when it will be possible to reveal eventual correlations between the distribution of the spins tilts and other parameters, we make the branching ratio between the preferentially aligned component and the isotropic component a function of the mass ratio. Specifically:

$$p(\cos\tau_1, \cos\tau_2 | q, f_{q=1}, \sigma_{c\tau}, \mu_{c\tau}, n) = \frac{1 - f_a(q, f_{q=1}, n)}{4} + f_a(q, f_{q=1}, n)\mathcal{N}(\cos\tau_1|\mu_{c\tau}, \sigma_{c\tau})\mathcal{N}(\cos\tau_2|\mu_{c\tau}, \sigma_{c\tau}), \quad (1)$$

where $\mathcal{N}$ indicates a normal distribution. We have defined the mass-ratio-dependent fraction of preferentially aligned systems as

$$\begin{aligned} f_a(q, f_{q=1}, n) &\equiv f_{q=1}\frac{g(q,n) - g(0.1,n)}{g(1,n) - g(0.1,n)} \\ g(q,n) &\equiv \exp\left[(q-0.1)^n - 0.9^n\right] \end{aligned} \quad (2)$$

and set $f_{q=1} = 1, \sigma_{c\tau} = 1.15, \mu_{c\tau} = 1$ and $n = 2$.

This non-linear correlation results in a higher fraction of systems with preferentially aligned spins as the mass ratio of the system gets closer to unity, and is shown in Fig. 1. While this functional form is not meant to represent a correlation based on solid astrophysical grounds, it can at least qualitatively capture one possible scenario in which dynamical environments, where isotropic spins are expected, produces BBHs with unequal masses.

To better visualize how the population we simulate compares with the default LVK results from GWTC-3, in the top panel of Fig. 2 we show the resulting distribution of the cosine tilts (marginalized over the mass ratio), and the 90% credible interval from the LVK analysis for comparison. The bottom panel shows the distribution of cosine tilts conditional on the mass ratio for $q$ in 4 different intervals together with the GWTC-3 90% credible interval. We notice that because the mass-ratio distribution peaks at equal masses (see e.g. Fig. 10 of Ref. [40]), in practice most sources in our simulated population will be drawn from a $\cos\tau$ distribution with a value of the aligned fraction $f_a$ close to 1.

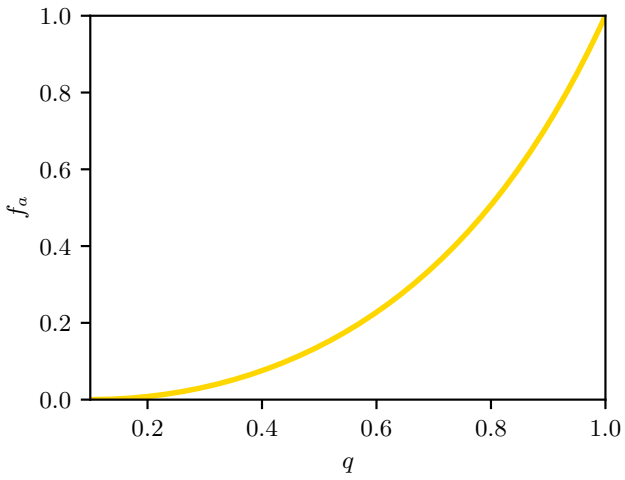



FIG. 1. Fraction of sources in the preferentially aligned-spin component as a function of the binary mass ratio. This distribution is used when generating the synthetic catalog of BBHs.

We generate a catalog of 1599 detectable BBHs[2] from this distribution, and add them into simulated LIGO-Virgo Gaussian noise, with a sensitivity corresponding to that of the fourth observing run, using the power spectral densities (PSD) provided by [74]. To mitigate the

[2] Ask me about this number next time you see me at a conference.

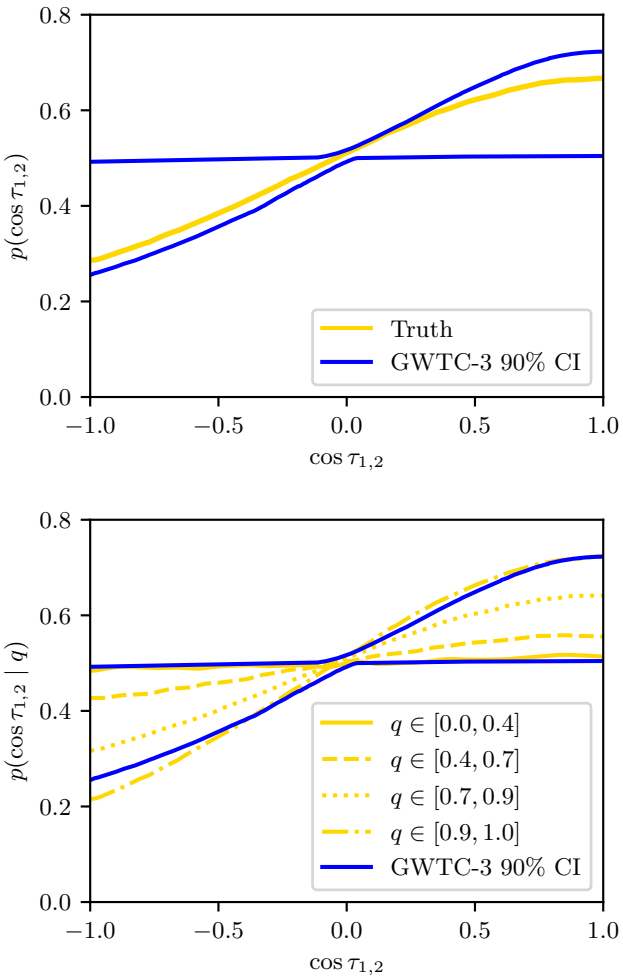



FIG. 2. The distribution of cosine tilts in the simulated universe (yellow), compared with the 90% credible interval from GWTC-3 (blue). The top panel shows the distributions for the two tilts (which are identical) marginalized over the mass ratio, while the bottom panel shows the conditional distribution for $\cos\tau_{1,2}$ for $q$ in four disjoint intervals.

problems highlighted by Ref. [75], we define detectability based on the matched filter signal-to-noise ratio (SNR), rather than the optimal SNR. Specifically, a source is detectable if its network matched-filter SNR — defined as the square root of the sum of the squares of the SNR of each detector — is greater than 11. We run the parameter estimation algorithm `Bilby` [76, 77] on the detectable sources, using the `IMRPhenomXP` [78] waveform model and the relative binning likelihood [79]. We do not include higher-order modes to keep the computation cost manageable (See discussion in Sec. VI). These sources are used to form catalogs of various size and analyzed using `GWPopulation` [80] to measure the hyper parameters, using the models described in Sec. IV.

## III. HIERARCHICAL LIKELIHOOD AND SELECTION EFFECTS

Bayesian inference[3] on catalogs of GW sources can be performed given a model for the underlying astrophysical population, parametrized by hyper parameters $\Lambda$ [81–85]:

$$p(\Lambda|D) \propto \pi(\Lambda) \prod_{i=1}^{N_{\rm events}} \frac{p(d_i|\Lambda)}{\alpha(\Lambda)}, \tag{3}$$

where $\alpha(\Lambda)$ is the population-dependent fraction of detectable events (see Eq. 8); $\pi(\Lambda)$ the hyper prior; $D = \{d_1, \ldots, d_{N_{\rm events}}\}$ is the data in the catalog; and $p(d_i|\Lambda)$ are the likelihoods of the individual events. This expression assumes that the merger rate has already been marginalized over using a log-uniform prior.

The choice of the population model affects the individual-event likelihoods which, upon explicit marginalization of the individual-event source parameters $\theta$, can be written as [86]:

$$\begin{aligned} p(d_i|\Lambda) &= \int \mathrm{d}\theta\, p(d_i|\theta)\pi(\theta|\Lambda) \\ &\propto \int \mathrm{d}\theta\, p(\theta|d_i, \mathrm{PE}) \frac{\pi(\theta|\Lambda)}{\pi(\theta|\mathrm{PE})}. \end{aligned} \tag{4}$$

The parameter estimation (PE) label indicates that the posterior $p(\theta|d_i, \mathrm{PE})$ is evaluated using a parameter estimation software, with associated prior $\pi(\theta|\mathrm{PE})$. We notice that this integral is formally inconsistent with our data selection procedure [75], since we condition detectability on the true values of each event source parameters (see also the discussion around Eq. (30) of Ref. [26]) but the systematics this approximation introduces are smaller than statistical uncertainties, see App. B. In practice, the integral in Eq. 4 is evaluated numerically as a Monte Carlo integral:

$$p(d_i|\Lambda) \simeq \frac{1}{M} \sum_{j=1}^{M} \frac{\pi(\theta_j|\Lambda)}{\pi(\theta_j|\mathrm{PE})}\bigg|_{\theta_j \sim p(\theta|d_i,\mathrm{PE})}. \tag{5}$$

The precision of this estimation depends on the number of posterior samples that are available. We run `Bilby` with the `Dynesty` [87] sampler and 4000 live points. This results in at least 16,000 posterior samples for each source.

In order to properly account for selection effects, the analyst must be able to calculate $\alpha(\Lambda)$, the fraction of detectable sources (with a definition of detectability that

[3] The reader familiar with data analysis for gravitational-wave populations can skip this section (though they might find Sec. III A interesting).

is self-consistent with that used to select the catalog) for all plausible values of the model hyper parameters. Specifically,

$$\alpha(\Lambda) = \int \mathrm{d}\theta \; p(\rho_\uparrow|\theta)\pi(\theta|\Lambda), \tag{6}$$

where the first term is the probability of a source with parameters $\theta$ to have detection statistics above threshold. This can be written as:

$$p(\rho_\uparrow|\theta) = \int \mathrm{d}D_\uparrow \; p(D_\uparrow|\theta), \tag{7}$$

where the integration domain is the data (i.e. noise realization) that would result on the source with parameters $\theta$ to be detectable. Plugging this expression into the previous one we have:

$$\alpha(\Lambda) = \iint \mathrm{d}\theta \; \mathrm{d}D_\uparrow \; \pi(\theta|\Lambda)p(D_\uparrow|\theta), \tag{8}$$

which can be evaluated by introducing a sampling distribution $\zeta$ from which the values of $\theta$ can be drawn:

$$\begin{aligned}\alpha(\Lambda) &= \iint \mathrm{d}\theta \; \mathrm{d}D_\uparrow \; \frac{\pi(\theta|\Lambda)}{\zeta(\theta|\mathcal{H}_\zeta)} p(D_\uparrow|\theta)\zeta(\theta|\mathcal{H}_\zeta) \\ &= \iint \mathrm{d}\theta \; \mathrm{d}D_\uparrow \; \frac{\pi(\theta|\Lambda)}{\zeta(\theta|\mathcal{H}_\zeta)} p(D_\uparrow\theta|\mathcal{H}_\zeta) \\ &= \frac{1}{N_{\rm tot}} \sum_{i=1}^{N_{\rm detectable}} \left.\frac{\pi(\theta_i|\Lambda)}{\zeta(\theta_i|\mathcal{H}_\zeta)}\right|_{\theta_i\sim\zeta(\theta|\mathcal{H}_\zeta)}, \end{aligned} \tag{9}$$

where we have numerically evaluated the integrals over data and $\theta$ by sampling values of $\theta$ from the sampling distribution $\zeta(\theta|\mathcal{H}_\zeta)$ and sampling values of the data (i.e. adding the signal associated to $\theta$ to a random segment of noise, real or simulated) and only keeping sources for which the resulting data stream corresponds to a detectable source [88]. $N_{\rm tot}$ sources will have to be drawn from $\zeta(\theta|\mathcal{H}_\zeta)$ before $N_{\rm detectable}$ are collected and in general $N_{\rm tot} \gg N_{\rm detectable}$, with the exact ratio depending on the details of the proposal distribution (e.g. the maximum redshift at which sources can be, and the mass function). A list of detectable sources to be used for evaluating the sum above during the analysis can be prepared in advance and stored to disk.

In general, the approximated estimate of $\alpha(\Lambda)$ will be better with more detectable sources [89, 90], and with a proposed distribution that is similar to the (unknown) astrophysical population. We create a catalog of detectable sources in two steps. First, we generate ∼8.3 millions detectable sources drawing their parameters from the same population model used to generate the sources in the catalog, i.e., $\zeta(\theta|\mathcal{H}_\zeta) = p(\theta|\Lambda_{\rm true})$. Then, we generate ∼2.4 millions detectable sources drawing their parameters from a population model that has the same mass and redshift distribution as the previous one, but has uniform spin magnitude and cosine tilt distribution. The two sets are then combined using the method described by Ref. [91] and the overall list of 10.7 million sources are used to evaluate $\alpha(\Lambda)$ numerically. In order to obtain $\mathcal{O}(10^7)$ detectable sources we have to generate $\mathcal{O}(10^9)$ sources, i.e., around 1% of sources are detectable. Unless otherwise stated, we run `GWPopulation` with a maximum total variance of 2 for the log likelihood [90], i.e., when sampling the hyper parameter space, samples that corresponds to a variance larger than 2 are rejected. In practice, for most of the runs the actual variance is much lower, and we will enforce a more stringent cut as discussed below.

We will often show the PPD of the astrophysical parameters that characterize individual sources, i.e., spins, masses and redshift. This can be thought of as the expected distribution of those parameters in light of the detected sources. It can be written as:

$$\begin{aligned} p(\theta|D) &= \int \mathrm{d}\Lambda \; \pi(\theta|\Lambda)p(\Lambda|D) \\ &= \frac{1}{M}\sum_{i=1}^{M} \pi(\theta|\Lambda_i)|_{\Lambda_i\sim p(\Lambda|D)} \end{aligned} \tag{10}$$

That is, the PPD is the expectation of the population model calculated with draws from the posterior $p(\Lambda|D)$. To make the notation lighter, we'll often use PPD($\theta$) to mean $p(\theta|D)$.

### A. A note about efficient production of detectable sources

We notice that producing a list of detectable signals for the evaluation of $\alpha(\Lambda)$ can be computationally very expensive, as one in general has to calculate the SNR of all sources, even those that won't eventually end up being detectable. This problem is even more severe when estimating the sensitivity of real GW searches, as one needs to also run search algorithms, not just calculate SNRs [39, 92, 93]. The problem becomes more important if one is sampling many low-mass and/or high-redshift sources, a smaller fraction of which are detectable. For our purposes, the process can be trivially parallelized over as many CPU cores as possible. To further enhance the efficiency of the algorithm, we also pre-build a look-up table as follows:

- For a given source-frame total mass, we assume equal masses, and take both spins magnitudes equal to 0.99 and aligned with the angular momentum. We generate a waveform with these intrinsic parameters.

- For each detector in the network, we calculate the optimal SNR that the source would have if it were

overhead and face-on. We calculate the square root of the squared sum of these SNRs. We calculate the redshift at which this SNR would be equal to 11.

We notice that – given our setup – both steps are conservative: a) for a given total mass, a system with equal masses produces the strongest signal (if higher order modes are neglected, as in our waveforms). Similarly, maximal spins aligned with the binary orbital angular momentum yield higher SNRs; b) a face-on source produces the strongest signal and obviously a source cannot be overhead to all detectors in the network. Thus, with both steps we are *overestimating* the optimal SNR that the source would produce[4].

With this method, we obtain a mapping $z_{\rm max}(M_{\rm tot})$ that can be quickly interpolated anywhere in the mass range where our population model is defined. This gives a very conservative estimate of the maximum redshift at which a source with given masses could be detectable. Therefore, when we sample $\zeta(\theta|\mathcal{H}_\zeta)$, we can compare the proposed redshift with $z_{\rm max}(M_{\rm tot})$ and only compute the waveform if the proposed redshift is below the maximum. For this waveform, a random noise realization is generated and a matched filter SNR calculated to decide if the source is detectable or not, in a way that is consistent with the way we selected the detectable sources of the simulated population we analyze. With our population, this precomputed look-up table with an over-conservative value of the horizon reduces to 20% (of $\mathcal{O}(10^9)$ draws) the number of sources for which a waveform must be generated. We have not attempted to optimize this process, and it is certainly possible that a further speed-up could be obtained.

### B. A note on variances

In this work we will consider catalogs comprising up to 1500 sources and a variety of models. It is well known that the total variance on the hyper log likelihood estimator depends on the catalog size [89, 90]. As mentioned above, samples with log likelihood variance larger than 2 are rejected already during sampling. In practice, with our settings, the total variance is usually smaller than 1 for all but the largest catalogs, comprising 1500 sources. In Fig. 3 we show the distribution of variances for the models described in Sec. IV and three exemplary catalog sizes. Each histogram represent the distribution of variances across hyper samples for one run. For the smaller catalogs, variances are usually as small as $\sim 0.1$, though we do occasionally have hyper samples that yield much larger variances. We treat those as outliers — samples that explore region of the parameter spaces where our numerical estimation of the likelihood is unusually uncertain and should not be trusted — and remove them when reporting numerical results and plots henceforth, unless otherwise stated. Specifically, we use the following thresholds for the various catalog sizes, informed by the highest 90% percentile of the total variance across models for any given catalog size:

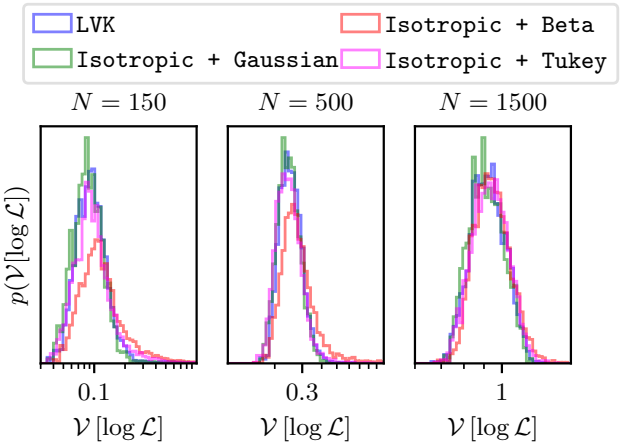



FIG. 3. Distribution of total variances for all models (described in Sec. IV) and catalog sizes (we do not show . The variances for correlated models look visually identical.

- $N = 69$, $\mathcal{V}[\log\mathcal{L}] < 0.3$
- $N = 150$, $\mathcal{V}[\log\mathcal{L}] < 0.75$
- $N = 300$, $\mathcal{V}[\log\mathcal{L}] < 1.0$
- $N = 500$, $\mathcal{V}[\log\mathcal{L}] < 1.0$
- $N = 1500$, $\mathcal{V}[\log\mathcal{L}] < 1.1$

## IV. HYPER MODELS

In this paper we will use a subset of $\cos\tau$ models from Ref. [64]:

[4] To decide whether any given source is detectable or not, we use the *matched filter* SNR, as mentioned above. The use of the optimal SNR while building this lookup table is not self-consistent, but also not consequential because our setup is so conservative that in practice we are only excluding sources that have an SNR well below our threshold of 11. Given that matched filter SNR and optimal SNR rarely differ by more than $\sim 1$, and often less, this inconsistency is not problematic. An indirect proof that this is indeed the case, is the fact that our redshift PPD— which is very sensitive to eventual issues with the estimation of the detectors' sensitivity—is not biased; see App. B.

$$\texttt{LVK}: \quad p(\cos\tau_1,\cos\tau_2|\sigma_{c\tau},f_a)=\frac{1-f_a}{4}+f_a\prod_{i=1}^{2}\mathcal{N}(\cos\tau_i,\mu_{c\tau}=1,\sigma_{c\tau}) \tag{11}$$

$$\texttt{Isotropic + Gaussian}: \quad p(\cos\tau_1,\cos\tau_2|\mu_{c\tau},\sigma_{c\tau},f_a)=\frac{1-f_a}{4}+f_a\prod_{i=1}^{2}\mathcal{N}(\cos\tau_i,\mu_{c\tau},\sigma_{c\tau}) \tag{12}$$

$$\texttt{Isotropic + Beta}: \quad p(\cos\tau_1,\cos\tau_2|\alpha_{c\tau},\beta_{c\tau},f_a)=\frac{1-f_a}{4}+f_a\prod_{i=1}^{2}\mathcal{B}(\cos\tau_i,\alpha_{c\tau},\beta_{c\tau}) \tag{13}$$

$$\texttt{Isotropic + Tukey}: \quad p(\cos\tau_1,\cos\tau_2|T_{x0},T_k,T_r,f_a)=\frac{1-f_a}{4}+f_a\prod_{i=1}^{2}\mathcal{T}(\cos\tau_i,T_{x0},T_k,T_r). \tag{14}$$

where $\mathcal{B}$ and $\mathcal{T}$ indicate a beta distribution and Tukey window, respectively, as defined in Eq. E.1 of Ref. [64]. For primary mass, mass ratio, spin magnitude and redshift, we use the same parametric models employed when generating the sources. We set hyper priors equal to Tab. G2 of Ref. [64] (the location of the `Isotropic + Gaussian`'s Gaussian component is restricted in the range $[-1,1]$), with the only difference that the two parameters controlling the beta component of `Isotropic + Beta` are uniform in the range $[0.05,7]$. These models can be run either assuming that the branching ratio $f_a$ is constant, or that it is correlated with the mass ratio, with the same functional form we used when simulating the sources, Eq. 2. For all models, the priors on $f_a$ is $\mathcal{U}[0,1]$ and – when enabling correlations – the prior on $n$ is $\mathcal{U}[0.01,12]$. When correlations are allowed we will append "corr" to label of the model (e.g., `Isotropic + Beta corr`)

We stress that the correlated `Isotropic + Gaussian` model can exactly match the true astrophysical distribution. This constitutes *an unrealistic best case scenario*, since in general for real data we won't have the luxury of believing that our parametric model is a perfect match to nature, for some values of its parameters. This specific analysis will thus represent the absolute best that can be done given the available data, when all of the possible sources of systematics, namely model mismatch, have been removed. The results obtained with it will constitute a useful quantitative assessment of what one might possible hope to obtain, in optimal conditions.

## V. RESULTS

### A. Is a peak at $\cos\tau \neq 1$ significative?

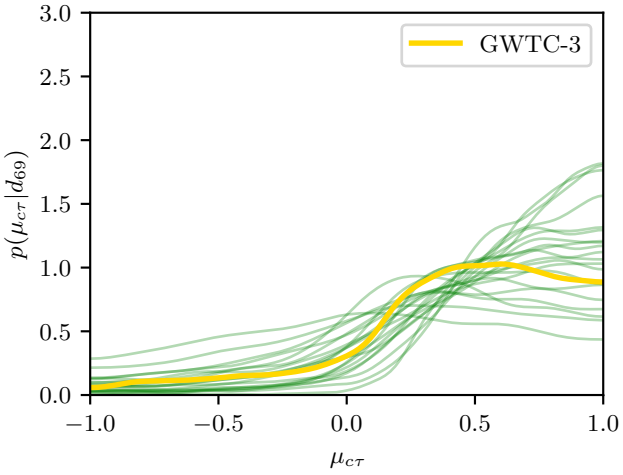
The broad peak or plateau found in the $\cos\tau$ distribution of the GWTC-3 data [21, 24, 25, 64] is surprising, as it is not obviously predicted or associated with any of the main astrophysical formation channels. At the same time, GWTC-3 only included 69 BBHs, and $\cos\tau$ is notoriously hard to measure, with most BBH sources producing very broad posteriors [70, 94].

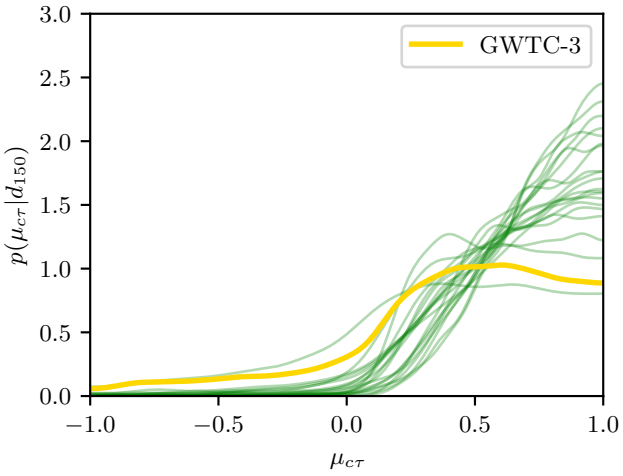
The first question we want to address is thus whether – given a limited catalog size – one could expect a peak in the PPD of $\cos\tau$ away from $\cos\tau = 1$ even if the spins in the true underlying distribution were in fact born that way. The answer will clearly depend on what the true distribution is and—for distributions that peak at $\cos\tau = 1$—exactly how broad the distribution is. Given that running these analyses is still relatively computationally expensive, we use our fiducial set of sources, described in Sec. II to attack this question[5]. To that end, we generated 20 catalogs of 69, 150, and 300 BBHs each, drawing from the 1500 sources for which we have produced single-event posteriors. We chose these three sizes to be somewhat representative of the sizes of the GWTC-3 catalog, and what might be available at the end of O4a and O4b. For each of these catalogs we perform the population analysis with the `Isotropic + Gaussian` model, without allowing for correlations [6]. In other words, we consider the minimal extension to the `LVK` model introduced in Vitale *et al.* [64].

In the top panel of Fig. 4 we show the hyper posterior measurements for $\mu_{c\tau}$ for each of the catalogs of 69 BBH (solid curves) and, for reference, the posterior obtained in Ref. [64] using the `Isotropic + Gaussian` on the GWTC-3 data (yellow curve). We find that—depending on the realization of events in each catalog—$\mu_{c\tau}$ can have rather different shapes: some catalogs yield relatively narrow distributions peaked at 1 (i.e. the true value) while others return more shallow distributions or plateau. The GWTC-3 hyper posterior is consistent with what we find here, and we conclude that it is not impossible that a true underlying distribution with a broad peak at $\cos\tau = 1$ might have yielded a measurement like GWTC-3's. We notice that for only 10% of our catalogs

[5] In Sec. VI we will comment on how these results could change given a different true distribution.

[6] We will show later that even much larger catalogs cannot definitively reveal such a correlation, and thus we are not significantly biasing the analyses in this section by not allowing for it.



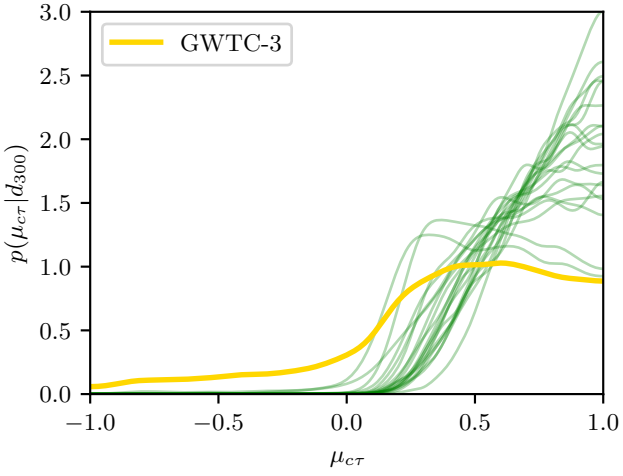




FIG. 4. Posteriors of $\mu_{c\tau}$ for each of 20 catalogs containing 69 (top panel), 150 (middle panel) and 300 (lower panel) sources. The yellow line shows the posterior obtained by Ref. [64] using the `Isotropic + Gaussian` model on GWTC-3.

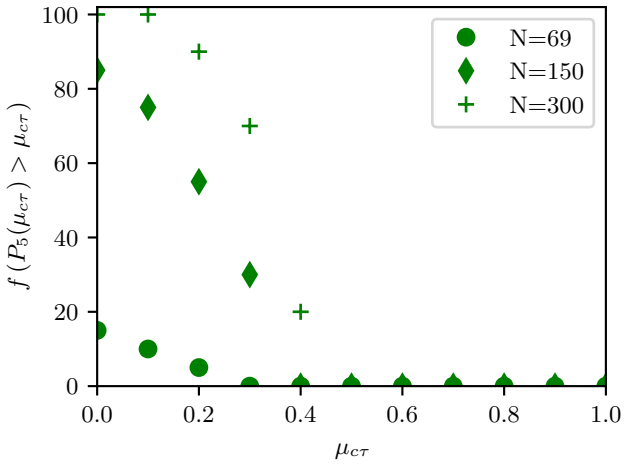



FIG. 5. For each value $\mu_{c\tau}$ in the abscissa, the fraction of catalogs for which the $5^{\rm th}$ percentile of $p(\mu_{c\tau}|d)$ is larger than $\mu_{c\tau}$. The catalog size is given in the legend and the average is taken over 20 catalogs for each size. The analyses are all performed with the `Isotropic + Gaussian` model.

of 69 BBH is the 5th percentile for $\mu_{c\tau}$ positive. This fraction increases to 85% (100%) with catalogs made of 150 (300) BBHs, as seen in the middle and bottom panels of Fig. 4.

In Fig. 5 we show for each of the catalog sizes the fraction of catalogs for which the $5^{\rm th}$ percentile of the $\mu_{c\tau}$ posterior is above the value given in the abscissa. We see that even for our larger catalogs there is a rather sudden drop of this fraction for abscissas in the range $[0, 0.4]$. For none of our simulated catalogs can we significantly constrain $\mu_{c\tau}$ to be above 0.5.

Even with larger catalogs comprising 300 sources it is not impossible or even unusual to find posteriors of $\mu_{c\tau}$ that peak far from 1, in some cases with peaks in the range $0 \leq \mu_{c\tau} \leq 0.5$. These posteriors in $\mu_{c\tau}$ are usually paired with broad posteriors for $\sigma_{c\tau}$, such that the resulting inference in $\cos\tau$ still supports a wide range of possibilities. This is a common issue with looking at individual, fully marginalized, posteriors of complicated models with many parameters. We usually still indulge in that exercise because it is often the case that some of these parameters have a clear and useful physical or astrophysical interpretation, and are directly linked to what the analysis is trying to measure.

Ultimately, what is being constrained is the high-dimensional shape of the model's parameters, which we can plot through the PPD of the relevant part of the model. It would be unpractical to show 20 sets of $\cos\tau$ PPD for each of the three catalog sizes we have considered in this section. Instead, for each size we chose three exemplary catalogs: one that yields a rather flat posterior, one that yields a peak away from unity, and one that peaks at unity. These are shown in Fig. 6: in each panel, the green dot-dashed line is the median and the

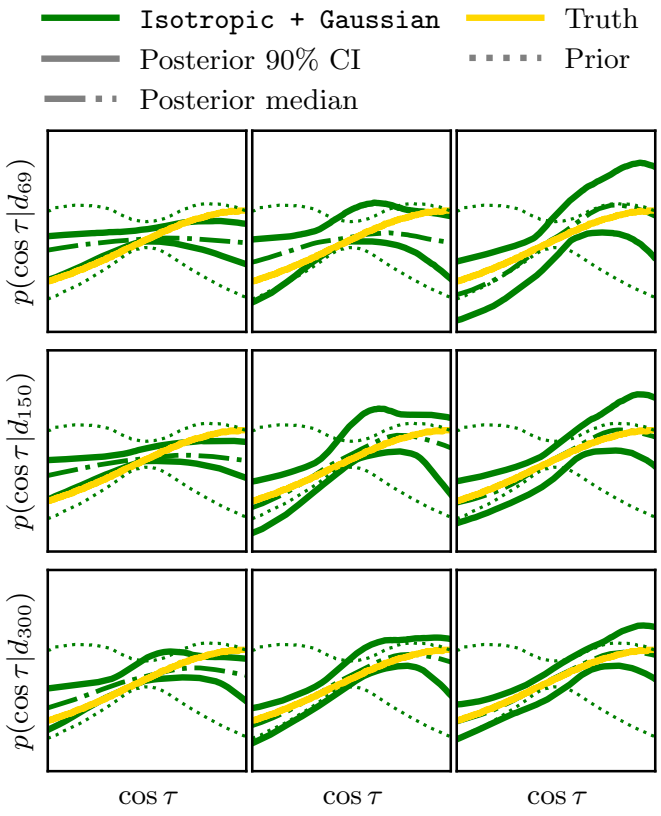



FIG. 6. For each of the catalog sizes, $N = 69$ (top row) $N = 150$ (middle row) and $N = 300$ (bottom row) three exemplary PPDs for $\cos\tau$. In each panel the yellow curve shows the truth, the cyan solid line the median, and the dark band the 90% credible interval. For the $N = 300$ catalogs, one can still get either broad 90% posterior plateaus (bottom left) or even peaks away from unity (bottom mid) though those are more rare than in smaller catalog sizes.

green solid lines enclose the 90% central credible region. Dotted lines enclose the central 90% of the prior and the solid yellow line shows the true distribution of $\cos\tau$. We see that it is not impossible for even a catalog with 300 BBHs to produce a flat PPD in $\cos\tau$, like the one we show in the bottom left panel. However, it is much more common for catalogs of that size to result in PPDs that peak toward unity, like the bottom right panel.

We thus conclude that, to the extent that the true distribution of $\cos\tau$ in nature is similar to what simulated here, it would not be surprising to measure spurious peaks in the PPD distribution away from unity after 150 BBHs are collected, and it would not be impossible (though more unlikely) after 300 BBHs are collected.

### B. The long road to alignment

In this section we report on the measurements we obtain with three catalogs, with $N = 150$, $N = 500$ and $N = 1500$ sources. Our data release also includes cat-

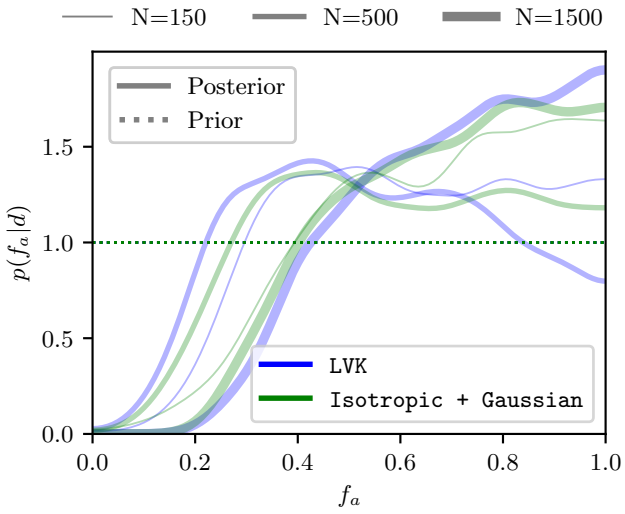



FIG. 7. Fraction of sources in the preferentially aligned component for the `LVK` and `Isotropic + Gaussian` models. The prior is marked with a dotted line (in this case, the same for both models).

alogs with $N = 300$ and $N = 1000$ sources but we are not showing these in the body of the paper to keep plots less busy. Given the current median merger rate form BBHs, this implies we are able to make projections that span the fourth and the fifth observing run[7]. We'll first focus on the results one obtains using models that do *not* allow for the $\cos\tau$–$q$ correlation we have introduced in our simulated universe in Sec. V B 1 (`LVK` and `Isotropic + Gaussian` models) and V B 2 (`Isotropic + Beta` and `Isotropic + Tukey` models). We allow for that correlation in Sec. V B 3.

#### 1. *Uncorrelated models: `LVK` and `Isotropic + Gaussian`*

**Hyper parameters**

The only thing that all models have in common, in addition to yielding a PPD for the same set of single-event parameters (masses, spins, redshift), is that they all include the fraction of sources in the non-isotropic component - $f_a$- as one of their parameters. Therefore, this parameter can always be compared across models and catalog sizes. We show the posterior on $f_a$ for the `LVK` and `Isotropic + Gaussian` models in Fig. 7. The plot shows that the two models perform similarly well, and that while the uncertainty shrinks as the number of

[7] Even though we used power spectral density representative of O4, our results can be used to make projections for O5 because the main impact of a more sensitive network is to increase the number of detectable sources.

sources increases, even the largest catalogs can't exclude small values of $f_a$. While even with the smallest catalog size we find that $f_a = 0$ is excluded at high credibility, values as small as $f_a \gtrsim 0.3$ cannot be ruled out even after 1500 sources.

It may be surprising that the uncertainty for $f_a$ shrinks relatively slowly as the number of events increases. In fact, for all the hyper parameters that control the spin tilt distributions it is *not* the case that the uncertainties scale with the square root of the number of events. The reason why this happens can be better understood by looking at a corner plot of all parameters that control the tilt distribution: $f_a$, $\mu_{c\tau}$ and $\sigma_{c\tau}$, which is shown in Fig. 8. As more and more events are added to the catalog, the main effect is that the joint $f_a$-$\sigma_{c\tau}$ projection shrinks along its semi-minor axis. This mainly helps exclude configurations with a large fraction of aligned sources (large $f_a$) and a narrow Gaussian peak (small $\sigma_{c\tau}$) centered close to $\mu_{c\tau} = 1$. Once that is done, it is harder to further exclude parts of the parameter space because—given the very uncertain measurement of $\cos\tau$ on an event by event basis—the analysis cannot easily differentiate between a universe with the true value of $\sigma_{c\tau}$ and another universe with a smaller value of $\sigma_{c\tau}$ which also produces slightly fewer sources with preferentially aligned spins[8]. We are left to contend with a correlated, high-dimensional parameter space oddity, where different combinations of the hyper parameters controlling the spin tilt distribution result in similar likelihoods. As we show below, quantities based on PPD should be preferred over marginalized hyper parameters.

For the `Isotropic + Gaussian` model, the parameter $\mu_{c\tau}$ has a clear astrophysical interpretation as the location of the peak of the non-isotropic component of the $\cos\tau$ distribution. Because of this direct interpretation, we quote uncertainties for it, the caveats about marginal posteriors notwithstanding. The top left panel of Fig. 8 shows the marginalized $\mu_{c\tau}$ posteriors for the catalogs with 150 and 1500 sources. We see that while in both cases the posterior peaks at 1, it is rather wide and smaller values are not excluded. We measure $\mu_{c\tau} = 0.74^{+0.24}_{-0.42}$, $\mu_{c\tau} = 0.72^{+0.25}_{-0.41}$ and $\mu_{c\tau} = 0.74^{+0.13}_{-0.30}$ for $N = 150$, $N = 500$ (not shown in the corner plot) and $N = 1500$ sources, respectively.

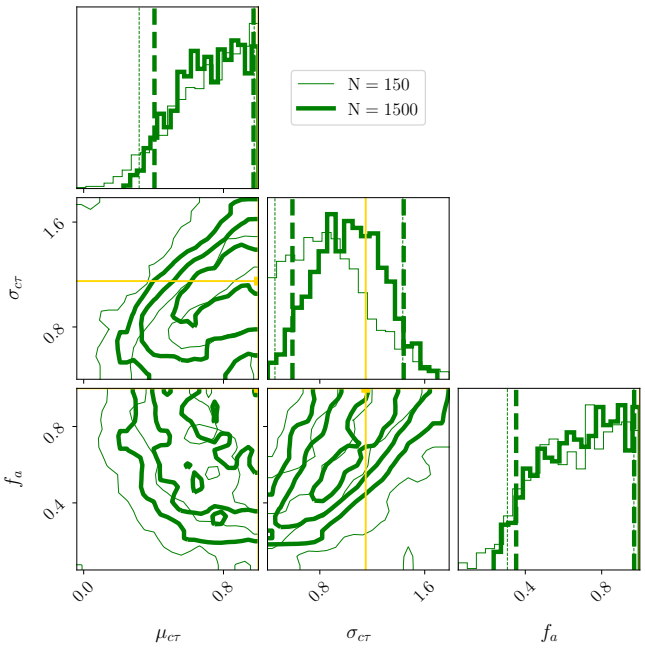



FIG. 8. Parameters that control the $\cos\tau$ part of the `Isotropic + Gaussian` model. To keep the corner plot lighter we only plot results for $N = 150$ and $N = 1500$. The yellow line shows the true values (note: there is no true value for $f_a$ since the true distribution has a $q$-dependent fraction of aligned events, Fig. 1). The contours in the 2D plots enclose 68.3%, 95.4%, and 99.7% of the posterior volume. The vertical dashed lines in the diagonal panels enclose 90% of the posterior. Priors are uniform for all parameters.

### PPD and derived quantities

In Fig. 9 we show the 90% credible intervals of the PPDs for $\cos\tau$ (PPDs for the other parameters are shown in App. B) obtained using the `LVK` and `Isotropic + Gaussian` models, with the yellow curve indicating the true distribution and dotted curves enclosing 90% of the prior. For each catalog size, the truth is included in the 90% credible interval everywhere in the domain of $\cos\tau$ for both models. The main difference we observe is that for the two larger catalog sizes the `Isotropic + Gaussian` model yields a flatter distribution for positive $\cos\tau$.

We can more easily compare the width of the uncertainty regions by plotting slices of the PPD at fixed values of $\cos\tau$. For example, we show a slice of the PPD $\cos\tau = 1$ for the catalog including 1500 sources in Fig. 10. Both models do well and, as one might have imagined, the simpler model with fewer parameters (and with $\mu_{c\tau} = 1$ by construction) yields a narrower PPD. The tail on the left of the `Isotropic + Gaussian` PPD corresponds to those flatter posteriors characteristic of the measurements obtained with that model. Smaller catalogs have correspondingly larger uncertainties: for the `Isotropic + Gaussian` model at $\cos\tau = 1$ the 90% uncertainty is 0.16 with 1500 sources, which becomes 0.21 (0.36) for catalogs of 500 (150) sources. This is roughly a twofold reduction in the uncertainty at $\cos\tau = 1$ as the number of sources increases tenfold. Slicing at different values, we find shrinkage between 2 and 3 compared to the catalog with 150 sources. Similar considerations can be made for the `LVK` model, for which the shrinkage is larger, between 3 and 4 going from 150 to 1500 sources, depending on where one slices. We quote uncertainties on the PPD sliced at other values in Tabs. II and III in App. C.

We can recast our PPD into an integrated estimation of

[8] Similar patterns can be found in the `LVK` model.

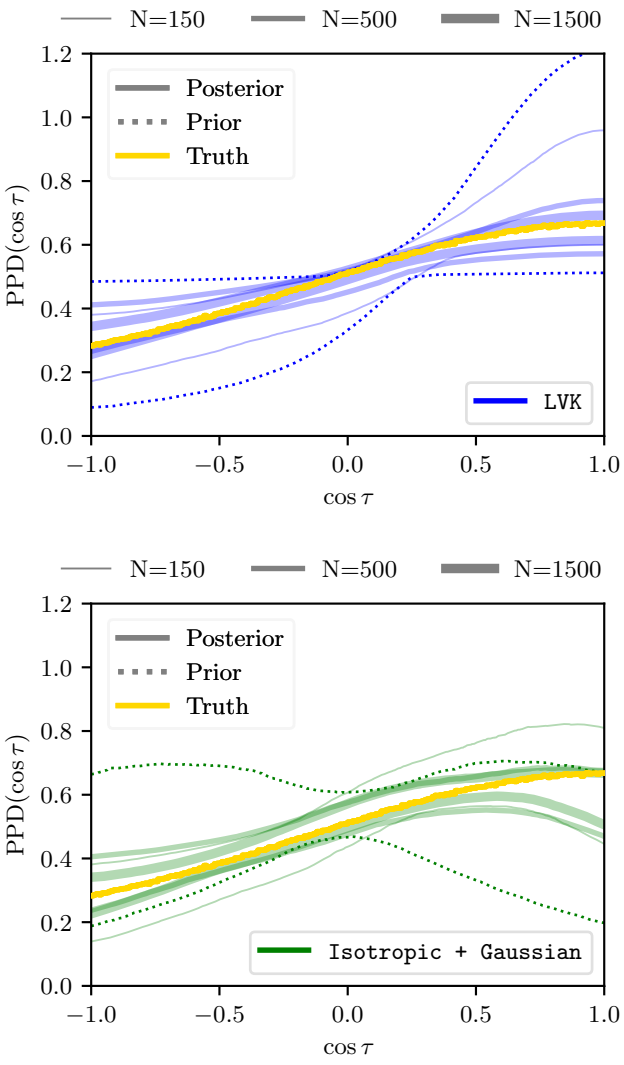



FIG. 9. PPD for $\cos\tau$ for catalogs of 150, 500, 1500 events using the uncorrelated `LVK` model (top) and `Isotropic + Gaussian` model (bottom).The thickness of the curves indicates the catalog size and each set of curves spans the 90% credible interval. The yellow curve is the true distribution.

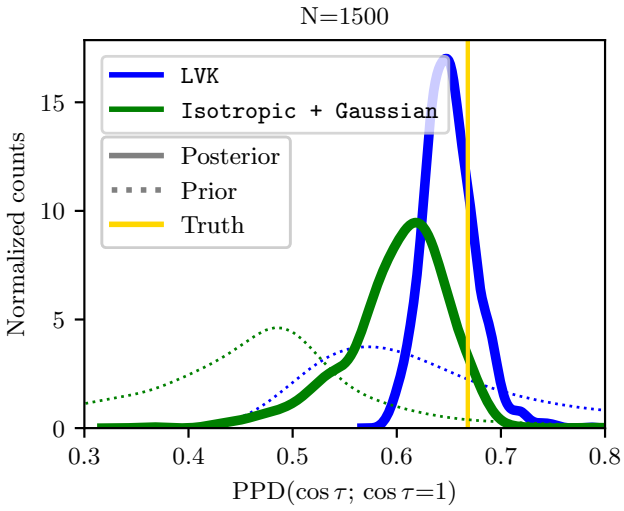



FIG. 10. Slice of the PPD for $\cos\tau$ at $\cos\tau = 1$ for the catalog with 1500 sources. The truth is indicated with a yellow line.

the fraction of tilts in the population that have $\cos\tau$ below (or above) any threshold. In Fig 11 we show the fraction of BHs with $\cos\tau \leq 0$ (top) and $\cos\tau \gtrsim 0.98$ (bottom) – that is, spin vectors closer than 10° relative to perfect alignment – for the `LVK` and `Isotropic + Gaussian` models, with the yellow lines marking the true values. We find that both models do well for both bounds. The true fraction of systems with negative tilts in the population is 38.9%; using the `LVK` model we find $36.5^{+6.8}_{-7.9}$%, $41.4^{+3.7}_{-4.2}$% and $40.0^{+2.8}_{-1.6}$% with $N = 150$, 500, and 1500 events, respectively. The `Isotropic + Gaussian` model yields $37.1^{+7.4}_{-7.8}$%, $41.8^{+3.9}_{-4.6}$%, $40.7^{+2.8}_{-2.2}$%. The true fraction of sources with $\cos\tau \gtrsim 0.98$ is 1.0%; using the `LVK` model we find $1.1^{+0.4}_{-0.2}$%, $1.0^{+0.2}_{-0.1}$% and $1.0^{+0.1}_{-0.1}$% with $N = 150$, 500 and 1500 events, respectively. The `Isotropic + Gaussian` model yields $1.0^{+0.3}_{-0.3}$%, $0.9^{+0.2}_{-0.2}$%, $0.9^{+0.1}_{-0.1}$%. For both bounds we find that uncertainty shrinks from by a factor of $\sim 3-4$ as the number of events increases tenfold. Other values are tabulated in Tabs. IV and V in App. C.

### 2. *Uncorrelated models - `Isotropic + Beta` and `Isotropic + Tukey`*

We now shift focus to the two other uncorrelated models included in this work, where the non-isotropic component is either a (possibly singular) beta distribution or a Tukey window. Those models have been introduced by Ref. [64] and used to analyze GWTC-3 data. They are more generic than the `LVK` model and its `Isotropic + Gaussian` extension in such that they can capture a more diverse set of morphologies, with non-Gaussian peaks or plateaus. We show the PPD for $\cos\tau$ obtained with both models in Fig. 12. At $\cos\tau \simeq 1$ the higher end of the 90% credible interval for the `Isotropic + Beta` model is out of scale due to some hyper samples that result in a singular beta distribution. The number of these samples decreases as the catalog size increases. For the `Isotropic + Tukey` model we find that the true is slightly outside of the 90% credible interval for $\cos\tau \gtrsim 0.7$, whilst it is included in the 90% credible interval for all other cases. For $N = 150$, the `Isotropic + Tukey` model shows a peak in the upper 90% credible interval, which is washed out as more events are added to the catalog.

The fact that the `Isotropic + Tukey` model slightly underestimates the value of the $\cos\tau$ PPD toward unity can be seen by plotting a slide of the PPD, as done for the `LVK` and `Isotropic + Gaussian` models in Fig. 10. We do this for the catalog comprising 1500 sources and both

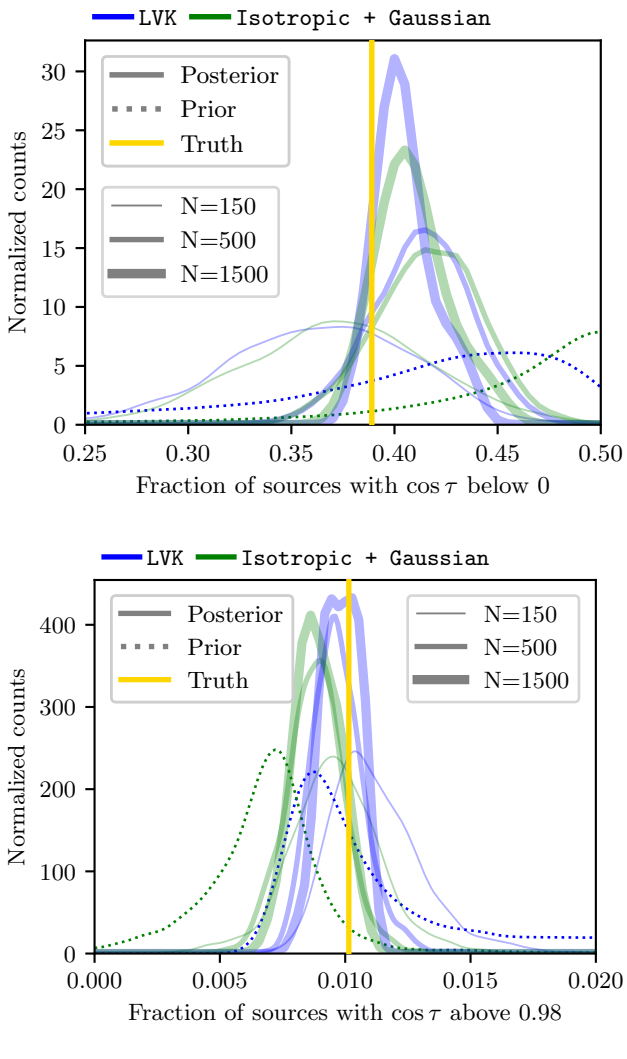



FIG. 11. (Top) derived posterior on the fraction of sources with $\cos\tau \leq 0$ for the `LVK` and `Isotropic + Gaussian` models and three catalog sizes. The true value is indicated with a yellow line. (Bottom) The same, but for the fraction of sources with $\cos\tau \gtrsim 0.98$ (i.e. within 10° from perfect alignment).

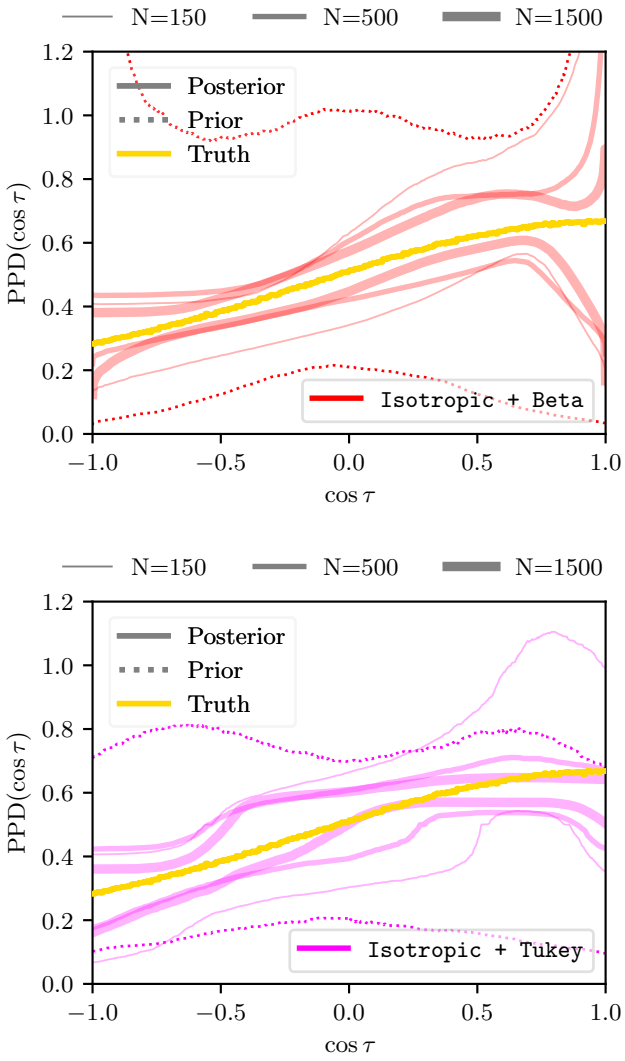



FIG. 12. PPD for $\cos\tau$ for catalogs of 150, 500, 1500 events using the uncorrelated `Isotropic + Beta` model (top) and `Isotropic + Tukey` model (bottom).The thickness of the curves indicates the catalog size and each set of curves spans the 90% credible interval. The yellow curve is the true distribution.

`Isotropic + Gaussian` and `Isotropic + Tukey` models in Fig. 13, where we actually slice at $\cos\tau = 0.99$ instead of 1.0 to avoid numerical issues with the large values that singular beta posteriors can take in the last bin. The true value is in the right-hand side of the sliced PPD for both models.

Relatedly, both models underestimate $f_a$, the fraction of tilts in the non-isotropic component. However, it's worth stressing again that because these two models are more elastic and do *not* enforce a Gaussian peak, the possible morphologies of the non-isotropic component are much richer, and include anything from a cosine peak, to a broad plateau or a skewed peak. This implies that the same value of $f_a$ across models can result in rather different PPDs for $\cos\tau$. This is another reason why is useful to work directly with quantities that are born off the PPD: they can be readily compared across models, irrespective of how they are parametrized (or even if they are non-parametric). We do this by looking again at the fraction of events with negative $\cos\tau$ and with $\cos\tau$ above $\sim 0.98$ for both models, Fig. 14. In most cases, the true value are included within the 90% credible intervals. Specifically, for the fraction of sources with $\cos\tau \leq 0$ the `Isotropic + Beta` (`Isotropic + Tukey`) model yields $35.0^{+8.0}_{-7.9}$%, $40.9^{+4.6}_{-5.0}$% and $40.2^{+1.7}_{-1.8}$% ($37.4^{+12.0}_{-9.8}$%, $41.8^{+4.6}_{-5.0}$% and $40.8^{+2.9}_{-2.1}$%) for $N = 150$, 500 and 1500 respectively, where the true value was 38.9%. The fraction of systems with $\cos\tau \gtrsim 0.98$ for the `Isotropic + Beta` (`Isotropic + Tukey`) model is measured to be $0.9^{+3.2}_{-0.6}$%, $0.7^{+1.5}_{-0.2}$%, $0.6^{+0.2}_{-0.2}$% ($0.9^{+0.6}_{-0.4}$%,

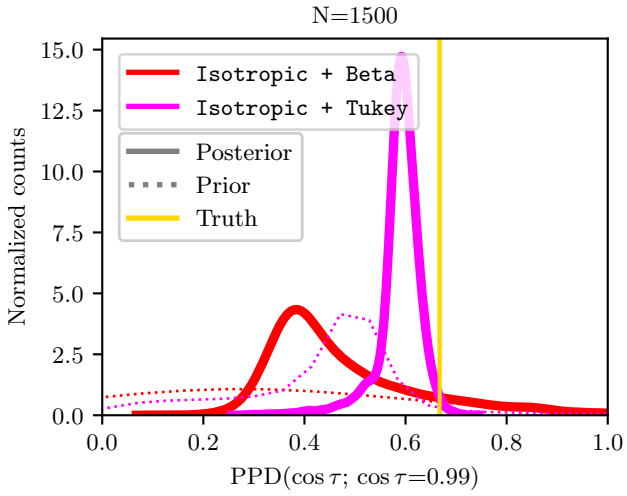



FIG. 13. Slice of the PPD for $\cos\tau$ at $\cos\tau = 0.99$ for the catalog with 1500 sources. The truth is indicated with a yellow line.

$0.9^{+0.2}_{-0.2}\%$, $0.9^{+0.1}_{-0.1}\%$) for $N = 150, 500, 1500$, respectively, where the true value was 1.0%.

### 3. *Correlated models*

In this section we focus on the results obtained when the branching fraction between the two components of the tilt models is allowed to vary with the mass ratio. This introduces another hyper parameter that we call $n$. The mass ratio-dependent branching fraction is thus parametrized by two quantities that are measured from the data, $f_{q=1}$ and $n$, which are combined as in Eq. 2 to give $f_a(q)$. Qualitatively speaking, $f_{q=1}$ is the fraction of sources in the non-isotropic component at $q = 1$ and $n$ controls how quickly the branching ratio evolves as $q$ increases, with $n = 0$ yielding a uniform branching ratio, and larger $n$ corresponding to a steeper increase of $f(q)$ as $q$ approaches 1.

We begin by anticipating that the results in this section are to a large extent the same as what presented in Secs. V B 1 and V B 2 because even for our largest catalogs it proves to be extremely challenging to measure this correlation. The main difference we observe is that the PPD for $\cos\tau$ obtained with the correlated models can be slightly more uncertain than what obtained with the corresponding non-correlated model even when the prior 90% region is smaller, i.e., favoring more uniform distributions. For example, in Fig. 15 we compare the PPD obtained with the `Isotropic + Gaussian corr` model with that of the `Isotropic + Gaussian`, for the catalog with 150 sources.

Therefore, in this section we will not show updated versions of the plots of Secs. V B 1 and V B 2 for the correlated models, as they will too look similar to justify

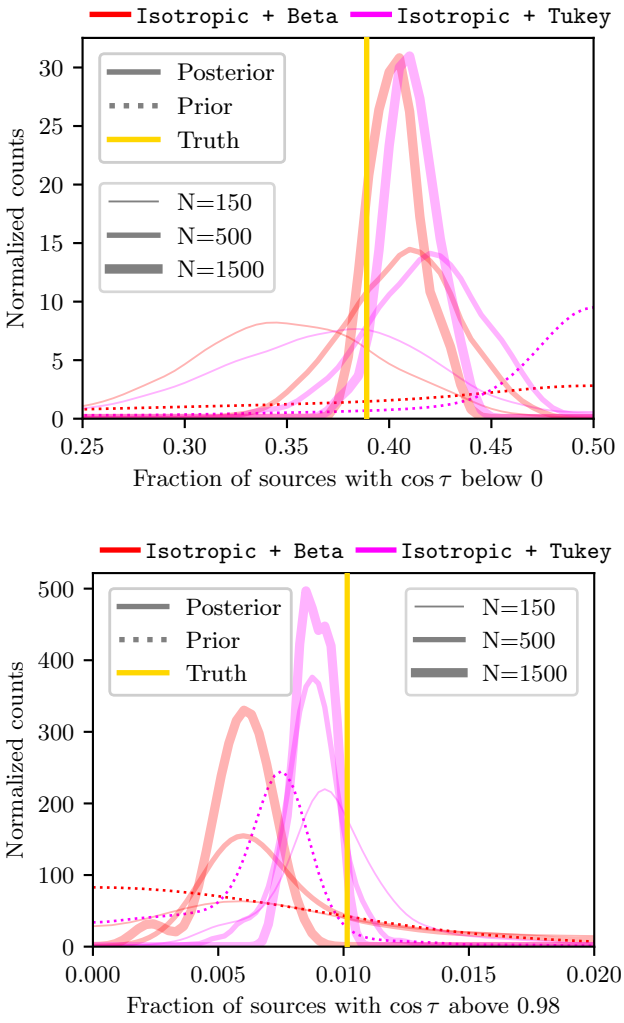



FIG. 14. Same as Fig. 11 but for the `Isotropic + Beta` and `Isotropic + Tukey` models.

the space they take. Instead, we will focus specifically on how well the mass ratio dependent branching fraction can be measured and whether the analysis can prove that there indeed is correlation in the data. In Fig. 16 we show the 90% credible interval for the measurement of $f(q)$ obtained using the `Isotropic + Gaussian corr` model (the `LVK corr` model yields virtually the same set of curves). The truth (yellow solid curve) is included in the 90% credible intervals except at the right edge, which is expected given that the 95$^{\text{th}}$ percentile must be smaller than the maximum, and the maximum cannot be larger than 1 for the branching ratio.

Three main things are worth stressing. First, the measurement is somewhat informative, meaning that we do not simply get the hyper prior back. That is shown in the figure as two dotted black lines enclosing 90% of the hyper prior. Second, the uncertainty bands are extremely wide, even for the largest catalog, highlighting the difficulty of measuring $f(q)$ even with a model

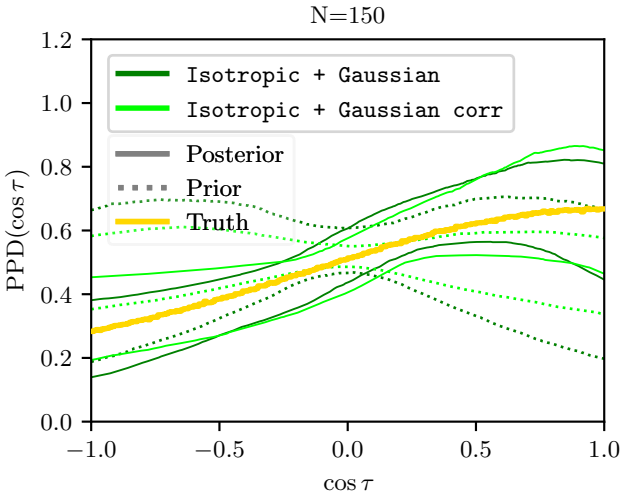



FIG. 15. PPD for $\cos\tau$ for catalogs of 150 events using the `Isotropic + Gaussian` model (green solid lines) and `Isotropic + Gaussian corr` model (lime solid lines). The yellow curve is the true distribution and the dotted lines are the priors.

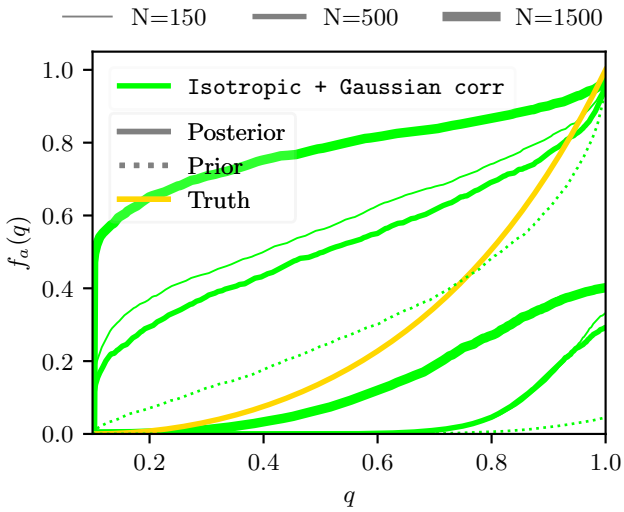



FIG. 16. Posterior on the mass ratio-dependent branching ratio $f(q)$ obtained with the `Isotropic + Gaussian corr` model. The green curves enclose the 90% credible regions while the yellow line is the true branching ratio and the black dotted lines enclose the 90% credible region of hyper prior draws.

that can perfectly match the truth. Third, the width of the uncertainty bands does not trivially shrink as the catalog size increases, especially at small $q$. We interpret this as a sign that not only the catalog size matters, but also which specific sources are included in the catalog. This can be understood because a good measurement of $f(q)$ at small $q$ requires the detection of sources with small $q$. Given that we are using a true mass-ratio distribution that favors equal-mass sources (compatibly with the LVK's measurement in GWTC-3) there are not that many small mass ratio sources in our catalogs to start with. This, together with the fact that $f_{q=1}$ and $n$ are heavily correlated with one another and partially correlated with the other spin magnitude and spin tilt parameters implies that the resulting posterior on $f(q)$ can appear to have this non-obvious progression with the number of sources.

To verify that nothing is wrong with the model itself (meaning that we are not affected by a bug), we have verified that running the population code with the `Isotropic + Gaussian corr` model fixing all parameters to their true values but either of $f_{q=1}$ or $n$ in turn, returns a 1D posterior that includes the corresponding true value. Even in the unrealistic scenario captured by this test (all hyper parameters known but one) the measurements are noisy. For example, when only $n$ is assumed unknown we get $n = 1.30^{+5.96}_{-1.17}$, $n = 5.00^{+6.10}_{-3.21}$ and $n = 2.15^{+1.73}_{-1.04}$ for $N = 150, 500$ and 1500 respectively. Only the largest catalog yields a 1D posterior that peaks at the true value, but still has a relative uncertainty of $\gtrsim 100\%$. The two remaining models yield the same qualitative results, with the `Isotropic + Beta corr` underestimating $f(q)$ for large $q$ in the smaller catalogs.

We end this section by focusing on the remaining question we wanted to address: can the analyst even claim that this $q - \cos\tau$ correlation exists in the detected population? Based on the fact that the PPDs of the correlated and uncorrelated models are so similar, it should not be surprising that the answer is not positive. Indeed, a constant, horizontal, $f(q)$ line can be drawn within the 90% credible intervals of Fig. 16 for any catalog size (the sharp decline on the left is due to the prior). In Tab. I we show for all our runs the natural log Bayes factor between each model and the "correct" model, i.e. the `LVK corr`. We find that for all catalog sizes the simpler `LVK` model that does not allow for the correlation is slightly preferred (though it should be remembered that the uncertainty on the Bayes factors we calculate with `Dynesty` with our settings is of the order of $\sim 0.4$). The model with the largest number of hyper parameters, `Isotropic + Tukey corr` is consistently ranked last. The table highlights that none of the models is definitively ruled out, and that it is not the case that the models allowing for the correlation are ranked higher.

## VI. CONCLUSIONS

In this paper, we looked into the prospects for measuring the astrophysical distribution of spin tilts (i.e., the angle between the spin vectors and the binary orbital angular momentum) for the stellar-mass black holes in the binaries detected by LIGO, Virgo and KAGRA. This is known to be a challenging problem, owing to the uncertain event-by-event measurements of this quan-

| Model | ln Bayes Factor |
|---|---|
| **150 events** | |
| `LVK` | +0.5 |
| `Isotropic + Gaussian` | −0.1 |
| `Isotropic + Beta` | −0.1 |
| `Isotropic + Beta corr` | −0.4 |
| `Isotropic + Tukey` | −0.5 |
| `Isotropic + Gaussian corr` | −0.7 |
| `Isotropic + Tukey corr` | −0.7 |
| **500 events** | |
| `LVK` | +0.6 |
| `Isotropic + Gaussian` | 0.0 |
| `Isotropic + Beta` | −0.4 |
| `Isotropic + Beta corr` | −0.4 |
| `Isotropic + Tukey` | −0.4 |
| `Isotropic + Gaussian corr` | −0.5 |
| `Isotropic + Tukey corr` | −0.8 |
| **1500 events** | |
| `LVK` | +1.1 |
| `Isotropic + Gaussian` | +0.5 |
| `Isotropic + Tukey` | 0.0 |
| `Isotropic + Beta` | −0.1 |
| `Isotropic + Gaussian corr` | −0.7 |
| `Isotropic + Tukey corr` | −1.0 |
| `Isotropic + Beta corr` | −1.2 |

TABLE I. For each group of runs with the same number of sources, the second column reports the natural log of the Bayes factor between the model given in the first column and the `LVK corr` model run with the corresponding number of events. A positive value favors the model in the row over the `LVK corr`. In each group, models are sorted by the Bayes factor.

tity. We have generated a synthetic population of 1599 BBH whose masses, redshifts and spins are drawn from a true underlying population that is consistent with what the LVK has measured in GWTC-3 data. Specifically, the spin tilts for both black holes are drawn from a distribution that has two components: one that produces isotropically distributed spin vectors and one that produces spins vectors that are drawn from a Gaussian distribution centered at $\cos\tau = 1$ with a width of 1.15. The branching ratio between these two components is mass-ratio dependent – Fig. 1 — representing an hypothetical scenario in which isolated binaries form with more equal masses while dynamically-formed binaries tend to have unequal masses. We used this large dataset to tackle four main questions.

**Is a peak at $\cos\tau \neq 1$ surprising?** - In Ref. [64] we showed that GWTC-3 is consistent with the astrophysical $\cos\tau$ distribution having a peak *away* from $\cos\tau = 1$, a result later corroborated by others. This is surprising since such a peak away from 1 does not have any known astrophysical interpretation. We have formed 20 catalogs with a random collection of 69 BBHs each, i.e., having the same size of GWTC-3 and analyzed them with the `Isotropic + Gaussian` model, where the mean of preferentially aligned $\cos\tau$ component is a free hyper parameter. We found that spurious peaks away from $\cos\tau = 1$ are not uncommon. We repeated this experiment with 20 catalogs of 150 or 300 sources, which are realistic catalog sizes at the end of O4a and O4b[9]. We find that, even for the largest catalog sizes, peaks away from 1 are not impossible, even though they become less common. Broad plateaus that span most of the positive $\cos\tau$ domain remain somewhat common. Indeed, it becomes much easier to at least constrain the mean of the preferentially aligned $\cos\tau$ component – $\mu_{c\tau}$– to be positive. That is possible for only 10% of the catalogs with 69 sources but with 100% for the catalogs with 300 sources. We conclude that the results based on GWTC-3 alone do not necessarily imply that the astrophysical $\cos\tau$ distribution does not have a peak at $\cos\tau = 1$.

*To the extent that our assumed population is representative of nature's, these results suggest that several hundred sources will be required before claims about the location of such peak, if one exists, can be made.*

**Can we measure the location of a peak?** - Next, we used our simulated sources to create even larger catalogs with $N = 150$, 500 or 1500 sources. Each of these was analyzed using 4 different spin tilt models (from Ref. [64]) that do *not* allow for any correlations. We found that measuring individual fully marginalized hyper parameters is challenging, due to the non-trivial correlations between them, and can depict a landscape grimmer than what we are actually facing. These includes the hyper parameters that control the fraction of non-isotropic spins, and the location of the non-isotropic component. In this sense, revealing the presence of a peak at $\cos\tau = 1$ like the one we simulated remains extremely challenging. We also looked at the PPD of the spin tilts. The `Isotropic + Gaussian` and `Isotropic + Tukey` models find very broad plateaus for the $\cos\tau$ distribution, spanning most of the positive $\cos\tau$ range even for the catalog with 1500 sources. The `Isotropic + Beta` model does find a peak at $\cos\tau \simeq 0.6$. However, all models yield PPDs consistent with the truth within 90% credible intervals for all catalog sizes with the exception of the `Isotropic + Tukey` model, that for $N = 1500$ slightly underestimates the true $\cos\tau$ distribution at $\cos\tau \gtrsim 0.6$. These results are partially driven by the fact that the true shape of the preferentially-aligned tilt component is quite broad (consistently with what found in GWTC-3).

*These findings highlight the inherent difficulty in conclusively identifying a peak at perfect alignment, unless the underlying astrophysical distribution is significantly narrower than the one we have modeled.*

[9] When we performed the bulk of these analysis, the LVK had not yet announced that there would be an O4c. [9]

**Can we measure precisely the fraction of events with negative tilt or nearly aligned spins?** - While finding the evidence of a peak might be elusive, all of the models we used did well in quantifying the fraction of systems with either negative tilts or with nearly aligned tilts. For example, using the `Isotropic + Gaussian` we measured the fraction of sources with negative $\cos\tau$ to be $37.1^{+7.4}_{-7.8}\%$ with the catalog comprising 150 sources. The true value for our population is 38.9% and calculating this quantity with prior draws[10] one gets for the `Isotropic + Gaussian` model $50.0^{+15.1}_{-15.8}$, thus this is truly a data-driven measurement. For our largest catalog, one gets $40.7^{+2.8}_{-2.2}\%$, that is a reduction of the uncertainties by a factor of $\sim 3$ with a tenfold increase in the number of sources. All models measure this parameter well, with the true value contained inside the 90% credible interval.

Perhaps even more interestingly, given the apparent difficulty in revealing peaks in the PPD for $\cos\tau$, is the measurement of the fraction of sources with tilts within a 10° angle from alignment (i.e., with $\cos\tau \gtrsim 0.98$). For our population, this is only 1% of black holes. For the `LVK` and `Isotropic + Gaussian` models the effective prior on this fraction peaks slightly below 1% (Fig. 11 bottom panel). For all catalog sizes the posterior shifts toward the truth, compared to the prior and shrinks as more sources are added. For the catalog with 1500 sources the `Isotropic + Gaussian` model measures this fraction to be $0.9^{+0.1}_{-0.1}\%$. The more general `Isotropic + Beta` and `Isotropic + Tukey` models do a similarly well for the fraction of negative tilts, with posteriors that include the truth and are significantly different from the respective priors. While the `Isotropic + Tukey` model does well even in measuring the fraction of sources with nearly aligned spin, the `Isotropic + Beta` model underestimate that fraction. The catalog with 150 sources yields a measurement that is barely different from the prior. The precision increases with the number of sources, at the expense of the accuracy: for the largest catalog the truth is just outside the 90% credible interval.

*Overall, these results suggest that integral quantities—such as the fraction of sources within a particular parameter range—are both less model dependent and have lower uncertainties than the marginal posteriors of model parameters, even when the two seem to have similar meanings.*

**Can we reveal a correlation between tilt angles and mass ratio?** - Given the large number of BBH sources simulated for this work, we thought it would be worth to not only assess the measurability of the spin tilt distribution but also whether an underlying astrophysical correlation can be revealed in the data. Given the GWTC-3 median merger rate estimate, even at O5 sensitivity it will take years to collect 1500 sources [95]. We therefore are probing what could be done toward the end of this decade. Sadly, the answer is disappointing. At least with the correlation and population we have assumed, we find that even the largest catalogs cannot measure a mass ratio-dependent branching fraction $f(q)$ that is definitively non-constant, i.e. correlated. The PPD on $\cos\tau$ obtained with models that allow for this correlation are not very different from the corresponding measurements performed with models that do not include such correlation. The fact that the data is not very informative about the correlation in our population was confirmed calculating Bayes factors between each model (with or without correlation) and the `LVK corr` model, which is the "true" one that was used to simulate the sources. We find that for all catalog sizes, the simpler `LVK` model that does not include correlations is favored, with natural log of the Bayes factor between 0.5 and 1.0 relative to the `LVK corr` model. All other models have either Bayes factors close to zero or are slightly disfavored relative to the `LVK corr` model. We notice that given our sampler settings, Bayes factors whose absolute values is smaller than $\sim 0.4$ should not be considered significant.

*These results indicate that confidently revealing subtle correlations between spin tilt angles and mass ratios—like the one we simulated—from GW observations will likely require substantially larger datasets, beyond what advanced detectors alone will provide this decade (but see caveats below).*

**Outlook** - Is it possible our choices for the simulated BBH population led to results there are too pessimistic? Here we lay down a series of caveats that might affect how the informed reader will think of this question. A peak at $\cos\tau = 1$ would likely be easier to reveal than what we reported here under a few scenarios: a) if there is a preferentially aligned component with a peak significantly narrower (e.g., a small $\sigma_{c\tau}$ in our `Isotropic + Gaussian` and `LVK` models) than what we simulated; b) if the true distribution of spin *magnitude* were to produce more high-spin sources than what simulated here, since the tilt of large spins is usually easier to measure than that of small spins [70]; c) if the true distribution of mass ratios were to produce more often unequal mass black holes than what assumed here, since the primary spins of black holes in large mass ratio systems are usually easier to measure [11]. Any combination of these three factors would likely improve the measurement of the $\cos\tau$ distribution, and in particular the existence of peak at nearly aligned spins.

However, it is important to admit that is possible that our simulated population is in fact more generous than nature's: we have assumed that systems with comparable masses (i.e. the majority) $f_{q=1} = 1$, while in reality

[10] For the simple `Isotropic + Gaussian` model, this number can actually be calculated analytically as $50^{+13}_{-13}\%$.

[11] This happens because for large mass ratios, the primary spin is more and more similar to $\chi_{\rm eff}$ [96–98] which is the best measured spin parameter [99]

the fraction of systems in the nearly-aligned component might be much smaller, making the measurement of the properties of that component more challenging. It is also entirely possible that the true spin magnitude distribution produces fewer systems with medium or large spin magnitudes, leaving the tilt measurements comfortably numb to subtle features.

It is also worth discussing caveats associated with inference models and waveforms. We included in our suites of models two that can perfectly well match the truth (`LVK corr` and `Isotropic + Gaussian corr`). We have done that to assess possible best-case scenarios that set a lower limit for the statistical and systematic uncertainties arising from the modeling of the populations. In reality, it is very unlikely any one of the parametric models currently used is a good representation of nature. Non-parametric models can be used, in either one or multiple dimensions, but the mitigated risk of systematics comes at the expense of increased statistical uncertainties, implying that even more sources than what we have considered would be required to reach a comparable level of statistical uncertainty. Regardless, it is noteworthy that, despite potential challenges from likelihood approximations [25, 89] and models' systematics [72] we are always able to correctly infer the underlying BBH distributions within uncertainties for all our models and parameters (App. B).

Everywhere in our analysis we used the same waveform family - to simulate the BBH added into synthetic noise, to analyze their properties and to evaluate the sensitivity of the detectors. This has removed the risk of systematic errors in a way that is probably not representative of what would happen when analyzing real data [71]. Additionally, we have not used waveforms with higher order modes, which could improve the measurements of some parameters, for heavy systems [100]. Furthermore, there is some evidence that numerical relativity surrogates [101] might yield more precise spin posteriors - at least for certain sources [94]. Should that be the case for enough sources, it might lead to more stringent constraints of the spin population parameters.

As advanced ground based detectors get more sensitive in the next few years, hundreds and then thousands of binary black holes will be detected. Next-generation detectors will reveal hundreds *of* thousands of these sources annually. We emphasize that accurate measurements of integrated properties of the astrophysical distribution of spin tilts are at reach already in the advanced detector era, though uncertainties are likely to remain larger than 10%. We encourage the population synthesis community to produce and quote these integrated numbers (e.g., the fraction of tilts above or below certain thresholds), wherever possible. Obtaining percent-level measurements or revealing more subtle features such as correlation between tilts and other parameters might have to wait for next-generation detectors. Eventually, under pressure from larger datasets, the balance will tilt toward a more detailed understanding of the origins of merging black hole binaries. One fine day we shall see clearly the details, as well as the big picture.

## ACKNOWLEDGMENTS

S.V. is partially supported by NSF through the grant PHY-2045740. M.M. is supported by the LIGO Laboratory through the National Science Foundation awards PHY-1764464 and PHY-2309200. The authors are grateful for computational resources provided by subMIT at MIT Physics and the LIGO Laboratory supported by National Science Foundation Grants PHY-0757058 and PHY-0823459. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. We thank the Institute for Nuclear Theory at the University of Washington for its kind hospitality and stimulating research environment. This research was supported in part by the INT's U.S. Department of Energy grant No. DE-FG02- 00ER41132. We thank Sofia Alvarez Lopez, Cristopher Berry, Sylvia Biscoveanu, Tom Callister, Tom Dent, Jack Heinzel, Simona Miller, Cailin Plunkett, Colm Talbot and Noah Wolfe for useful comments and feedback on this work. We thanks Davide Gerosa and our other colleagues in the SPINS: our spins might be non-aligned but our interests are.

The hyper posteriors samples produced for this work will be available in Zenodo [102] at the time of publication.

S.V. would like to dedicate this work to the memory of Aldo Francesco Vitale.

[1] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. Lett. **116**, 061102 (2016), arXiv:1602.03837 [gr-qc].
[2] J. Aasi *et al.* (LIGO Scientific), Class. Quant. Grav. **32**, 074001 (2015), arXiv:1411.4547 [gr-qc].
[3] F. Acernese *et al.* (VIRGO), Class. Quant. Grav. **32**, 024001 (2015), arXiv:1408.3978 [gr-qc].
[4] Y. Aso, Y. Michimura, K. Somiya, M. Ando, O. Miyakawa, T. Sekiguchi, D. Tatsumi, and H. Yamamoto (KAGRA), Phys. Rev. D **88**, 043007 (2013), arXiv:1306.6747 [gr-qc].
[5] R. Abbott *et al.* (KAGRA, VIRGO, LIGO Scientific), Phys. Rev. X **13**, 041039 (2023), arXiv:2111.03606 [gr-qc].
[6] A. H. Nitz, C. D. Capano, S. Kumar, Y.-F. Wang, S. Kastha, M. Schäfer, R. Dhurkunde, and M. Cabero, Astrophys. J. **922**, 76 (2021), arXiv:2105.09151 [astro-

ph.HE].
[7] S. Olsen, T. Venumadhav, J. Mushkin, J. Roulet, B. Zackay, and M. Zaldarriaga, Phys. Rev. D **106**, 043009 (2022), arXiv:2201.02252 [astro-ph.HE].
[8] LIGO, Virgo, and KAGRA Collaborations, Public alerts during the O4 observing run, `https://gracedb.ligo.org/superevents/public/O4/` (2024).
[9] LIGO, Virgo, and KAGRA Collaborations, LIGO-Virgo-KAGRA Observing Plans, `https://observing.docs.ligo.org/plan/` (2025).
[10] M. Mapelli, Formation Channels of Single and Binary Stellar-Mass Black Holes (2021) arXiv:2106.00699 [astro-ph.HE].
[11] T. A. Callister, Arxiv (2024), arXiv:2410.19145 [astro-ph.HE].
[12] M. Favata, C. Kim, K. G. Arun, J. Kim, and H. W. Lee, Phys. Rev. D **105**, 023003 (2022), arXiv:2108.05861 [gr-qc].
[13] K. Belczynski, V. Kalogera, F. A. Rasio, R. E. Taam, A. Zezas, T. Bulik, T. J. Maccarone, and N. Ivanova, Astrophys. J. Suppl. **174**, 223 (2008), arXiv:astro-ph/0511811.
[14] J. J. Eldridge, E. R. Stanway, L. Xiao, L. A. S. McClelland, G. Taylor, M. Ng, S. M. L. Greis, and J. C. Bray, Publications of the Astronomical Society of Australia **34**, e058 (2017), arXiv:1710.02154 [astro-ph.SR].
[15] T. Fragos, J. J. Andrews, S. S. Bavera, C. P. L. Berry, S. Coughlin, A. Dotter, P. Giri, V. Kalogera, A. Katsaggelos, K. Kovlakas, S. Lalvani, D. Misra, P. M. Srivastava, Y. Qin, K. A. Rocha, J. Román-Garza, J. G. Serra, P. Stahle, M. Sun, X. Teng, G. Trajcevski, N. H. Tran, Z. Xing, E. Zapartas, and M. Zevin, Astrophysical Journal Supplement Series **264**, 45 (2023), arXiv:2202.05892 [astro-ph.SR].
[16] J. Riley *et al.* (COMPAS Team, Team COMPAS), Astrophys. J. Supp. **258**, 34 (2022), arXiv:2109.10352 [astro-ph.IM].
[17] M. Spera, M. Mapelli, N. Giacobbo, A. A. Trani, A. Bressan, and G. Costa, Mon. Not. Roy. Astron. Soc. **485**, 889 (2019), arXiv:1809.04605 [astro-ph.HE].
[18] C. L. Rodriguez, N. C. Weatherford, S. C. Coughlin, P. Amaro-Seoane, K. Breivik, S. Chatterjee, G. Fragione, F. Kıroğlu, K. Kremer, N. Z. Rui, C. S. Ye, M. Zevin, and F. A. Rasio, Astrophysical Journal Supplement Series **258**, 22 (2022), arXiv:2106.02643 [astro-ph.GA].
[19] S. Vitale, W. M. Farr, K. Ng, and C. L. Rodriguez, Astrophys. J. Lett. **886**, L1 (2019), arXiv:1808.00901 [astro-ph.HE].
[20] A. Ray, I. Magaña Hernandez, S. Mohite, J. Creighton, and S. Kapadia, Astrophys. J. **957**, 37 (2023), arXiv:2304.08046 [gr-qc].
[21] T. A. Callister and W. M. Farr, Phys. Rev. X **14**, 021005 (2024), arXiv:2302.07289 [astro-ph.HE].
[22] A. M. Farah, T. A. Callister, J. M. Ezquiaga, M. Zevin, and D. E. Holz, Astrophys. J. **978**, 153 (2025), arXiv:2404.02210 [astro-ph.CO].
[23] B. Edelman, Z. Doctor, J. Godfrey, and B. Farr, Astrophys. J. **924**, 101 (2022), arXiv:2109.06137 [astro-ph.HE].
[24] B. Edelman, B. Farr, and Z. Doctor, Astrophys. J. **946**, 16 (2023), arXiv:2210.12834 [astro-ph.HE].
[25] J. Golomb and C. Talbot, Phys. Rev. D **108**, 103009 (2023), arXiv:2210.12287 [astro-ph.HE].
[26] J. Heinzel, M. Mould, and S. Vitale, Phys. Rev. D **111**, L061305 (2025), arXiv:2406.16844 [astro-ph.HE].
[27] I. Mandel, W. M. Farr, A. Colonna, S. Stevenson, P. Tiňo, and J. Veitch, Mon. Not. Roy. Astron. Soc. **465**, 3254 (2017), arXiv:1608.08223 [astro-ph.HE].
[28] J. Heinzel, M. Mould, S. Álvarez-López, and S. Vitale, Phys. Rev. D **111**, 063043 (2025), arXiv:2406.16813 [astro-ph.HE].
[29] A. Toubiana, M. L. Katz, and J. R. Gair, Mon. Not. Roy. Astron. Soc. **524**, 5844 (2023), arXiv:2305.08909 [gr-qc].
[30] V. Tiwari and S. Fairhurst, Astrophys. J. Lett. **913**, L19 (2021), arXiv:2011.04502 [astro-ph.HE].
[31] J. Godfrey, B. Edelman, and B. Farr, Arxiv (2023), arXiv:2304.01288 [astro-ph.HE].
[32] J. Sadiq, T. Dent, and M. Gieles, Astrophys. J. **960**, 65 (2024), arXiv:2307.12092 [astro-ph.HE].
[33] S. Vitale, R. Lynch, R. Sturani, and P. Graff, Class. Quant. Grav. **34**, 03LT01 (2017), arXiv:1503.04307 [gr-qc].
[34] C. Talbot and E. Thrane, Astrophys. J. **856**, 173 (2018), arXiv:1801.02699 [astro-ph.HE].
[35] C. Talbot and E. Thrane, Phys. Rev. D **96**, 023012 (2017), arXiv:1704.08370 [astro-ph.HE].
[36] M. Fishbach, D. E. Holz, and W. M. Farr, Astrophys. J. Lett. **863**, L41 (2018), arXiv:1805.10270 [astro-ph.HE].
[37] T. A. Callister, C.-J. Haster, K. K. Y. Ng, S. Vitale, and W. M. Farr, Astrophys. J. Lett. **922**, L5 (2021), arXiv:2106.00521 [astro-ph.HE].
[38] S. Biscoveanu, T. A. Callister, C.-J. Haster, K. K. Y. Ng, S. Vitale, and W. M. Farr, Astrophys. J. Lett. **932**, L19 (2022), arXiv:2204.01578 [astro-ph.HE].
[39] R. Abbott *et al.* (LIGO Scientific, Virgo), Astrophys. J. Lett. **913**, L7 (2021), arXiv:2010.14533 [astro-ph.HE].
[40] R. Abbott *et al.* (LIGO Scientific, VIRGO, KAGRA Scientific), arXiv e-prints , arXiv:2111.03634 (2021), arXiv:2111.03634 [astro-ph.HE].
[41] 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, Astrophys. J. **910**, 152 (2021), arXiv:2011.10057 [astro-ph.HE].
[42] K. W. K. Wong, K. Breivik, K. Kremer, and T. Callister, Phys. Rev. D **103**, 083021 (2021), arXiv:2011.03564 [astro-ph.HE].
[43] K. W. K. Wong and D. Gerosa, Phys. Rev. D **100**, 083015 (2019), arXiv:1909.06373 [astro-ph.HE].
[44] S. Colloms, C. P. L. Berry, J. Veitch, and M. Zevin, Arxiv (2025), arXiv:2503.03819 [astro-ph.HE].
[45] C. Plunkett, M. Mould, and S. Vitale, arxiv (2025), arXiv:2504.18615 [gr-qc].
[46] G. Franciolini, V. Baibhav, V. De Luca, K. K. Y. Ng, K. W. K. Wong, E. Berti, P. Pani, A. Riotto, and S. Vitale, Phys. Rev. D **105**, 083526 (2022), arXiv:2105.03349 [gr-qc].
[47] A. Q. Cheng, M. Zevin, and S. Vitale, Astrophys. J. **955**, 127 (2023), arXiv:2307.03129 [astro-ph.HE].
[48] J. Heinzel, S. Biscoveanu, and S. Vitale, Phys. Rev. D **109**, 103006 (2024), arXiv:2312.00993 [astro-ph.HE].
[49] D. Gerosa, M. Kesden, E. Berti, R. O'Shaughnessy, and U. Sperhake, Phys. Rev. D **87**, 104028 (2013), arXiv:1302.4442 [gr-qc].
[50] D. Gerosa, R. O'Shaughnessy, M. Kesden, E. Berti, and U. Sperhake, Phys. Rev. D **89**, 124025 (2014),

arXiv:1403.7147 [gr-qc].
[51] W. M. Farr, S. Stevenson, M. Coleman Miller, I. Mandel, B. Farr, and A. Vecchio, Nature **548**, 426 (2017), arXiv:1706.01385 [astro-ph.HE].
[52] M. Mould and D. Gerosa, Phys. Rev. D **105**, 024076 (2022), arXiv:2110.05507 [astro-ph.HE].
[53] V. Varma, S. Biscoveanu, M. Isi, W. M. Farr, and S. Vitale, Phys. Rev. Lett. **128**, 031101 (2022), arXiv:2107.09693 [astro-ph.HE].
[54] V. Varma, M. Isi, S. Biscoveanu, W. M. Farr, and S. Vitale, Phys. Rev. D **105**, 024045 (2022), arXiv:2107.09692 [astro-ph.HE].
[55] S. Biscoveanu, Arxiv (2025), arXiv:2502.04278 [astro-ph.HE].
[56] J. Roulet, H. S. Chia, S. Olsen, L. Dai, T. Venumadhav, B. Zackay, and M. Zaldarriaga, Phys. Rev. D **104**, 083010 (2021), arXiv:2105.10580 [astro-ph.HE].
[57] S. Stevenson, C. P. L. Berry, and I. Mandel, Mon. Not. Roy. Astron. Soc. **471**, 2801 (2017), arXiv:1703.06873 [astro-ph.HE].
[58] C. Périgois, M. Mapelli, F. Santoliquido, Y. Bouffanais, and R. Rufolo, Universe **9**, 507 (2023), arXiv:2301.01312 [astro-ph.HE].
[59] C. L. Rodriguez, M. Zevin, C. Pankow, V. Kalogera, and F. A. Rasio, Astrophys. J. Lett. **832**, L2 (2016), arXiv:1609.05916 [astro-ph.HE].
[60] P. Marchant, K. M. W. Pappas, M. Gallegos-Garcia, C. P. L. Berry, R. E. Taam, V. Kalogera, and P. Podsiadlowski, Astron. Astrophys. **650**, A107 (2021), arXiv:2103.09243 [astro-ph.SR].
[61] C. L. Fryer and V. Kalogera, Astrophys. J. **554**, 548 (2001), arXiv:astro-ph/9911312.
[62] V. Kalogera, Astrophys. J. **541**, 319 (2000), arXiv:astro-ph/9911417.
[63] C. L. Fryer, K. Belczynski, G. Wiktorowicz, M. Dominik, V. Kalogera, and D. E. Holz, Astrophysics Journal **749**, 91 (2012), arXiv:1110.1726 [astro-ph.SR].
[64] S. Vitale, S. Biscoveanu, and C. Talbot, Astron. Astrophys. **668**, L2 (2022), arXiv:2209.06978 [astro-ph.HE].
[65] Y.-J. Li, Y.-Z. Wang, S.-P. Tang, and Y.-Z. Fan, Phys. Rev. Lett. **133**, 051401 (2024), arXiv:2303.02973 [astro-ph.HE].
[66] M. V. van der Sluys, C. Röver, A. Stroeer, V. Raymond, I. Mandel, N. Christensen, V. Kalogera, R. Meyer, and A. Vecchio, Astrophys. J. Lett. **688**, L61 (2008), arXiv:0710.1897 [astro-ph].
[67] M. van der Sluys, I. Mandel, V. Raymond, V. Kalogera, C. Rover, and N. Christensen, Class. Quant. Grav. **26**, 204010 (2009), arXiv:0905.1323 [gr-qc].
[68] S. Vitale, R. Lynch, J. Veitch, V. Raymond, and R. Sturani, Phys. Rev. Lett. **112**, 251101 (2014), arXiv:1403.0129 [gr-qc].
[69] M. Pürrer, M. Hannam, and F. Ohme, Phys. Rev. D **93**, 084042 (2016), arXiv:1512.04955 [gr-qc].
[70] S. Vitale, R. Lynch, V. Raymond, R. Sturani, J. Veitch, and P. Graff, Phys. Rev. D **95**, 064053 (2017), arXiv:1611.01122 [gr-qc].
[71] A. Dhani, S. Völkel, A. Buonanno, H. Estelles, J. Gair, H. P. Pfeiffer, L. Pompili, and A. Toubiana, Arxiv (2024), arXiv:2404.05811 [gr-qc].
[72] S. J. Miller, Z. Ko, T. Callister, and K. Chatziioannou, Phys. Rev. D **109**, 104036 (2024), arXiv:2401.05613 [gr-qc].
[73] D. Wysocki, J. Lange, and R. O'Shaughnessy, Phys. Rev. D **100**, 043012 (2019).
[74] LIGO Scientific Collaboration, Noise curves used for Simulations in the update of the Observing Scenarios Paper, howpublished = `https://dcc.ligo.org/ligo-t2000012/public`, note = LIGO-T2000012-v5, year = 2020.
[75] R. Essick and M. Fishbach, Astrophys. J. **962**, 169 (2024), arXiv:2310.02017 [gr-qc].
[76] G. Ashton *et al.*, Astrophys. J. Suppl. **241**, 27 (2019), arXiv:1811.02042 [astro-ph.IM].
[77] I. M. Romero-Shaw *et al.*, Mon. Not. Roy. Astron. Soc. **499**, 3295 (2020), arXiv:2006.00714 [astro-ph.IM].
[78] G. Pratten, S. Husa, C. Garcia-Quiros, M. Colleoni, A. Ramos-Buades, H. Estelles, and R. Jaume, Phys. Rev. D **102**, 064001 (2020), arXiv:2001.11412 [gr-qc].
[79] K. Krishna, A. Vijaykumar, A. Ganguly, C. Talbot, S. Biscoveanu, R. N. George, N. Williams, and A. Zimmerman, Arxiv (2023), arXiv:2312.06009 [gr-qc].
[80] C. Talbot, A. Farah, S. Galaudage, J. Golomb, and H. Tong, Arxiv (2024), arXiv:2409.14143 [astro-ph.IM].
[81] T. J. Loredo and I. M. Wasserman, Astrophysical Journal Supplement Series **96**, 261 (1995).
[82] T. J. Loredo and I. M. Wasserman, Astrophys. J. **502**, 75 (1998), arXiv:astro-ph/9701111.
[83] T. J. Loredo and D. Q. Lamb, Phys. Rev. D **65**, 063002 (2002), arXiv:astro-ph/0107260.
[84] I. Mandel, W. M. Farr, and J. R. Gair, Mon. Not. Roy. Astron. Soc. **486**, 1086 (2019), arXiv:1809.02063 [physics.data-an].
[85] S. Vitale, D. Gerosa, W. M. Farr, and S. R. Taylor, Handbook of Gravitational Wave Astronomy "10.1007/978-981-15-4702-7_45-1" (2020), arXiv:2007.05579 [astro-ph.IM].
[86] E. Thrane and C. Talbot, Publications of the Astronomical Society of Australia **36**, e010 (2019), arXiv:1809.02293 [astro-ph.IM].
[87] J. S. Speagle, MNRAS **493**, 3132 (2020), https://academic.oup.com/mnras/article-pdf/493/3/3132/32890730/staa278.pdf.
[88] V. Tiwari, Class. Quant. Grav. **35**, 145009 (2018), arXiv:1712.00482 [astro-ph.HE].
[89] W. M. Farr, Research Notes of the American Astronomical Society **3**, 66 (2019), arXiv:1904.10879 [astro-ph.IM].
[90] C. Talbot and J. Golomb, Mon. Not. Roy. Astron. Soc. **526**, 3495 (2023), arXiv:2304.06138 [astro-ph.IM].
[91] R. Essick, Research Notes of the AAS **5**, 220 (2021).
[92] R. Essick, Phys. Rev. D **108**, 043011 (2023), arXiv:2307.02765 [gr-qc].
[93] A. Lorenzo-Medina and T. Dent, Class. Quant. Grav. **42**, 045008 (2025), arXiv:2408.13383 [gr-qc].
[94] T. Islam, A. Vajpeyi, F. H. Shaik, C.-J. Haster, V. Varma, S. E. Field, J. Lange, R. O'Shaughnessy, and R. Smith, Arxiv (2023), arXiv:2309.14473 [gr-qc].
[95] R. W. Kiendrebeogo *et al.*, Astrophys. J. **958**, 158 (2023), arXiv:2306.09234 [astro-ph.HE].
[96] T. Damour, Phys. Rev. D **64**, 124013 (2001), arXiv:gr-qc/0103018.
[97] E. Racine, Phys. Rev. D **78**, 044021 (2008), arXiv:0803.1820 [gr-qc].
[98] L. Santamaria *et al.*, Phys. Rev. D **82**, 064016 (2010), arXiv:1005.3306 [gr-qc].
[99] R. Abbott *et al.* (LIGO Scientific, VIRGO, KAGRA), arXiv e-prints , arXiv:2111.03606 (2021),

arXiv:2111.03606 [gr-qc].
[100] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, D. Gerosa, L. C. Stein, L. E. Kidder, and H. P. Pfeiffer, Phys. Rev. Research. **1**, 033015 (2019), arXiv:1905.09300 [gr-qc].
[101] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, L. E. Kidder, and H. P. Pfeiffer, Phys. Rev. D **99**, 064045 (2019), arXiv:1812.07865 [gr-qc].
[102] S. Vitale and M. Mould, Data for: The long road to alignment: Measuring black hole spin orientation with expanding gravitational-wave datasets (2025).

## Appendix A: Simulated population and GWTC-3 results

As mentioned in the body of the paper, the BBH population we simulated is consistent with that measured using the LVK model in GWTC-3 data [64]. This is shown in Fig. 17 below, where posteriors from GWTC-3 are shown in blue, and the yellow marks the values used for our simulations. Notice that the fraction of sources in the non-isotropic tilt component, $f_a$, is mass-ratio dependent in our population.

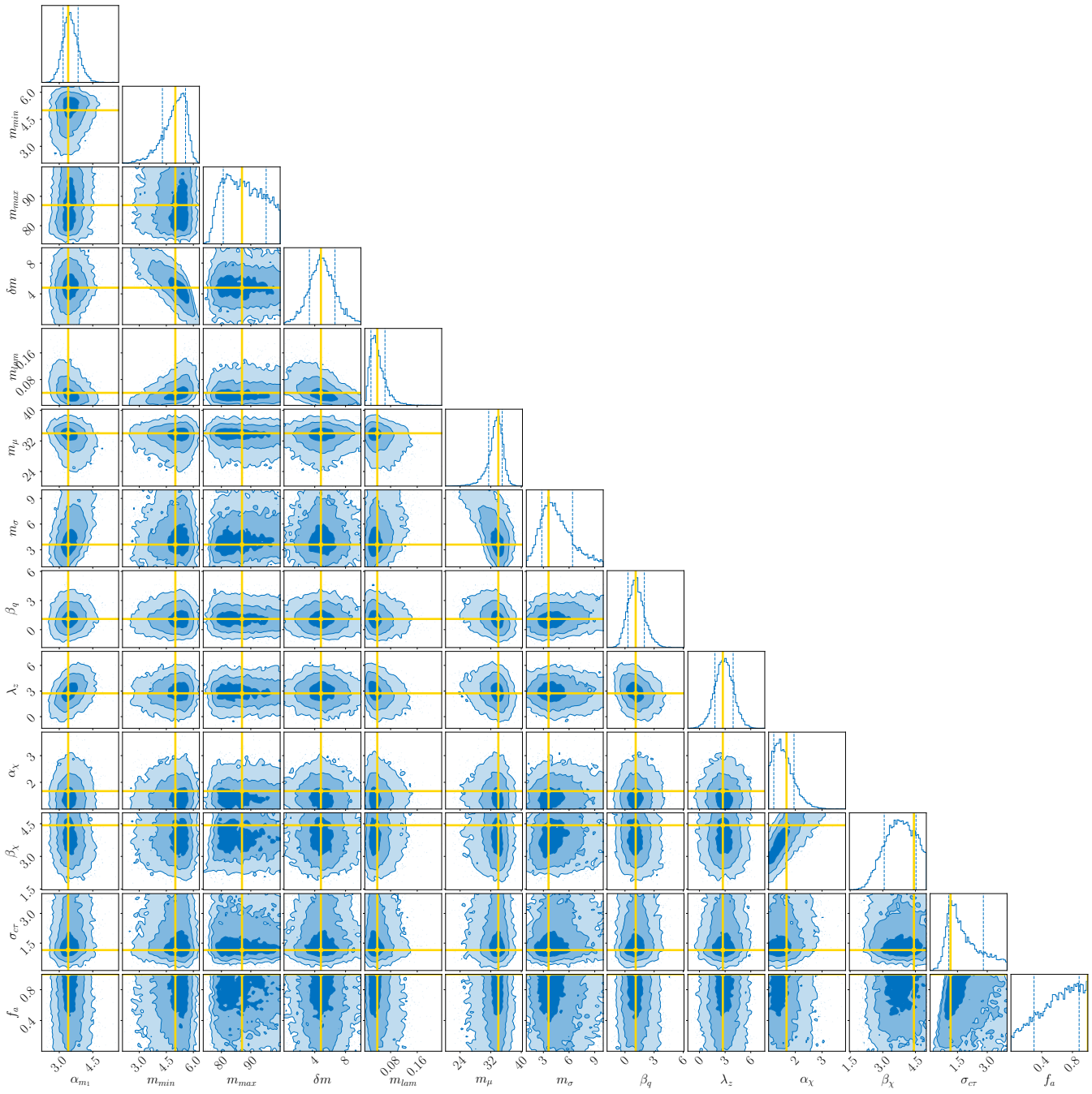



FIG. 17. The corner plot shows the GWTC-3 hyper posterior of the LVK model. The yellow lines mark the values used to generate our synthetic population.

### Appendix B: PPD for other hyper parameters

Even though the focus of this paper are the tilt angles, we should at least mention that the measurements of all other parameters – primary mass $m_1$, mass ratio $q \in [0.1, 1]$, redshift $z$ and spin magnitude $\chi$ – are also unbiased for all models and catalog sizes.

In Fig. 18 we show the PPDs for these parameters obtained with the `Isotropic + Gaussian` model and all three catalog sizes. The solid curves enclose 90% of the posterior and the dotted lines 90% of the prior. The yellow curves are the true distributions of these parameters and are generally included within the 90% credible intervals. No significant biases are seen anywhere. The PPDs for mass, mass ratio and redshift are nearly indistinguishable when our other models are used. For the `Isotropic + Beta` and `Isotropic + Beta corr` models only, a small difference is visible in $\chi$'s PPD: the upper edge of the credible region moves up by $\sim 10\%$ at $\chi \in [0 - 0.2]$ compared to the `Isotropic + Gaussian` and all other models.

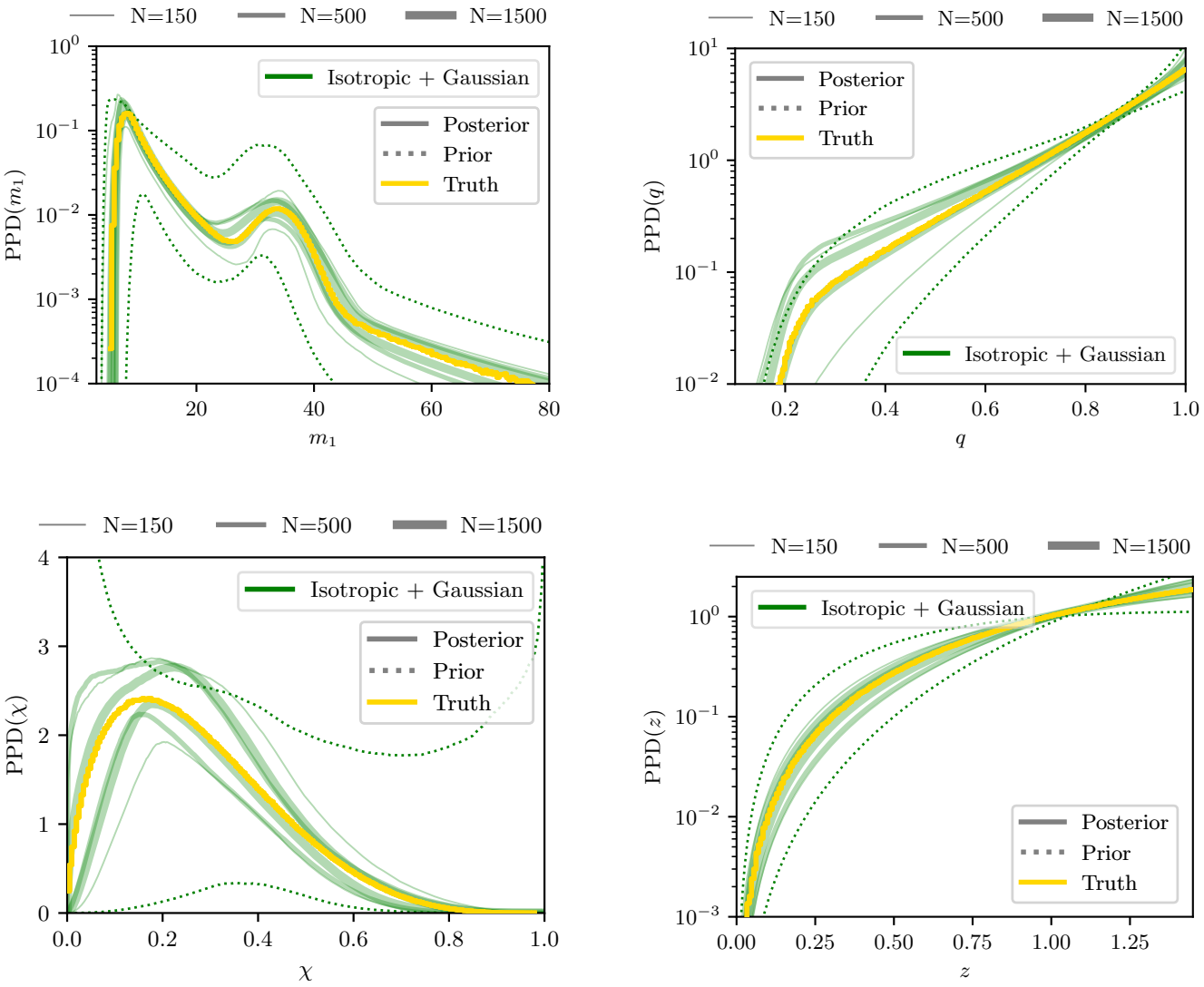



FIG. 18. PPD for the parameters $m_1$, $q$, $\chi$, and $z$ for the `Isotropic + Gaussian` model. Posteriors are shown for catalogs with 150, 500, and 1500 sources.

## Appendix C: Tables

We provide some tabulated values for median and 90% uncertainties for posterior and priors for the PPD sliced at $-1, -0.5, 0, 0.5, 1$ in Tab. II (uncorrelated models) and Tab. III (correlated models). Tab. IV and Tab. V report median and 90% uncertainties for posterior and priors for the fraction of sources with $\cos\tau \leq -0.98$, $\cos\tau \leq 0$ and $\cos\tau \geq 0.98$. The true values are given in all tables for every quantity.

| $\cos\tau$ | Truth | Model | 150 events | 500 events | 1500 events | Prior |
|---|---|---|---|---|---|---|
| -1 | 0.28 | `LVK` | $0.28^{+0.10}_{-0.11}$ | $0.34^{+0.08}_{-0.07}$ | $0.30^{+0.05}_{-0.04}$ | $0.34^{+0.14}_{-0.25}$ |
| | | `Isotropic + Gaussian` | $0.25^{+0.13}_{-0.12}$ | $0.32^{+0.08}_{-0.09}$ | $0.28^{+0.06}_{-0.06}$ | $0.47^{+0.20}_{-0.28}$ |
| | | `Isotropic + Beta` | $0.30^{+0.11}_{-0.18}$ | $0.36^{+0.07}_{-0.14}$ | $0.31^{+0.07}_{-0.19}$ | $0.29^{+7.61}_{-0.26}$ |
| | | `Isotropic + Tukey` | $0.27^{+0.14}_{-0.20}$ | $0.32^{+0.10}_{-0.15}$ | $0.26^{+0.10}_{-0.10}$ | $0.50^{+0.21}_{-0.40}$ |
| -0.5 | 0.39 | `LVK` | $0.36^{+0.07}_{-0.09}$ | $0.41^{+0.04}_{-0.05}$ | $0.40^{+0.02}_{-0.03}$ | $0.42^{+0.07}_{-0.27}$ |
| | | `Isotropic + Gaussian` | $0.36^{+0.08}_{-0.09}$ | $0.41^{+0.04}_{-0.05}$ | $0.39^{+0.03}_{-0.03}$ | $0.50^{+0.19}_{-0.18}$ |
| | | `Isotropic + Beta` | $0.34^{+0.09}_{-0.09}$ | $0.40^{+0.05}_{-0.06}$ | $0.38^{+0.04}_{-0.04}$ | $0.47^{+0.46}_{-0.34}$ |
| | | `Isotropic + Tukey` | $0.35^{+0.18}_{-0.13}$ | $0.41^{+0.11}_{-0.09}$ | $0.40^{+0.08}_{-0.08}$ | $0.50^{+0.30}_{-0.33}$ |
| 0 | 0.51 | `LVK` | $0.50^{+0.02}_{-0.12}$ | $0.51^{+0.01}_{-0.06}$ | $0.51^{+0.01}_{-0.02}$ | $0.50^{+0.02}_{-0.17}$ |
| | | `Isotropic + Gaussian` | $0.52^{+0.08}_{-0.09}$ | $0.52^{+0.05}_{-0.05}$ | $0.53^{+0.05}_{-0.03}$ | $0.52^{+0.09}_{-0.05}$ |
| | | `Isotropic + Beta` | $0.44^{+0.19}_{-0.10}$ | $0.48^{+0.12}_{-0.06}$ | $0.51^{+0.06}_{-0.06}$ | $0.52^{+0.50}_{-0.31}$ |
| | | `Isotropic + Tukey` | $0.55^{+0.11}_{-0.25}$ | $0.55^{+0.06}_{-0.16}$ | $0.57^{+0.04}_{-0.07}$ | $0.50^{+0.20}_{-0.30}$ |
| 0.5 | 0.62 | `LVK` | $0.65^{+0.09}_{-0.07}$ | $0.59^{+0.05}_{-0.04}$ | $0.61^{+0.03}_{-0.02}$ | $0.59^{+0.26}_{-0.08}$ |
| | | `Isotropic + Gaussian` | $0.65^{+0.11}_{-0.09}$ | $0.60^{+0.06}_{-0.05}$ | $0.62^{+0.03}_{-0.03}$ | $0.50^{+0.20}_{-0.17}$ |
| | | `Isotropic + Beta` | $0.65^{+0.21}_{-0.14}$ | $0.62^{+0.13}_{-0.10}$ | $0.65^{+0.09}_{-0.07}$ | $0.47^{+0.46}_{-0.34}$ |
| | | `Isotropic + Tukey` | $0.64^{+0.22}_{-0.19}$ | $0.59^{+0.10}_{-0.05}$ | $0.60^{+0.04}_{-0.03}$ | $0.50^{+0.29}_{-0.34}$ |
| 1 | 0.67 | `LVK` | $0.72^{+0.24}_{-0.12}$ | $0.63^{+0.11}_{-0.06}$ | $0.65^{+0.04}_{-0.04}$ | $0.62^{+0.59}_{-0.11}$ |
| | | `Isotropic + Gaussian` | $0.63^{+0.18}_{-0.19}$ | $0.58^{+0.10}_{-0.11}$ | $0.61^{+0.06}_{-0.10}$ | $0.47^{+0.20}_{-0.27}$ |
| | | `Isotropic + Beta`[(1)] | $0.67^{+1.63}_{-0.44}$ | $0.46^{+0.94}_{-0.16}$ | $0.43^{+0.38}_{-0.11}$ | $0.32^{+3.26}_{-0.29}$ |
| | | `Isotropic + Tukey` | $0.62^{+0.37}_{-0.27}$ | $0.58^{+0.10}_{-0.15}$ | $0.59^{+0.05}_{-0.09}$ | $0.50^{+0.18}_{-0.40}$ |

TABLE II. PPD (median and 90% credible interval) of $\cos\tau$ at fixed values of $\cos\tau$ (first column) for the uncorrelated models. The prior and true value are also reported. [(1)] For the `Isotropic + Beta` only, the slice is actually taken at $\cos\tau = 0.99$ numerical issues with the large values that singular beta posteriors can take in the last bin.

| $\cos\tau$ | Truth | Model | 150 events | 500 events | 1500 events | Prior |
|---|---|---|---|---|---|---|
| -1 | 0.28 | `LVK corr` | $0.33^{+0.10}_{-0.11}$ | $0.37^{+0.06}_{-0.08}$ | $0.30^{+0.05}_{-0.04}$ | $0.43^{+0.06}_{-0.15}$ |
| | | `Isotropic + Gaussian corr` | $0.32^{+0.14}_{-0.12}$ | $0.36^{+0.07}_{-0.09}$ | $0.29^{+0.06}_{-0.06}$ | $0.49^{+0.10}_{-0.13}$ |
| | | `Isotropic + Beta corr` | $0.33^{+0.09}_{-0.12}$ | $0.38^{+0.06}_{-0.09}$ | $0.31^{+0.07}_{-0.16}$ | $0.41^{+3.73}_{-0.19}$ |
| | | `Isotropic + Tukey corr` | $0.33^{+0.11}_{-0.15}$ | $0.37^{+0.07}_{-0.12}$ | $0.27^{+0.09}_{-0.09}$ | $0.50^{+0.09}_{-0.19}$ |
| -0.5 | 0.39 | `LVK corr` | $0.37^{+0.09}_{-0.09}$ | $0.41^{+0.05}_{-0.05}$ | $0.39^{+0.02}_{-0.03}$ | $0.46^{+0.04}_{-0.14}$ |
| | | `Isotropic + Gaussian corr` | $0.38^{+0.10}_{-0.11}$ | $0.41^{+0.05}_{-0.05}$ | $0.39^{+0.03}_{-0.03}$ | $0.50^{+0.10}_{-0.08}$ |
| | | `Isotropic + Beta corr` | $0.35^{+0.08}_{-0.09}$ | $0.40^{+0.05}_{-0.05}$ | $0.38^{+0.04}_{-0.04}$ | $0.49^{+0.22}_{-0.17}$ |
| | | `Isotropic + Tukey corr` | $0.36^{+0.14}_{-0.11}$ | $0.40^{+0.08}_{-0.07}$ | $0.39^{+0.10}_{-0.09}$ | $0.50^{+0.14}_{-0.15}$ |
| 0 | 0.51 | `LVK corr` | $0.48^{+0.04}_{-0.11}$ | $0.49^{+0.02}_{-0.07}$ | $0.51^{+0.01}_{-0.03}$ | $0.50^{+0.01}_{-0.09}$ |
| | | `Isotropic + Gaussian corr` | $0.51^{+0.07}_{-0.10}$ | $0.51^{+0.05}_{-0.06}$ | $0.52^{+0.05}_{-0.03}$ | $0.51^{+0.04}_{-0.02}$ |
| | | `Isotropic + Beta corr` | $0.43^{+0.15}_{-0.09}$ | $0.46^{+0.10}_{-0.05}$ | $0.51^{+0.07}_{-0.06}$ | $0.51^{+0.25}_{-0.14}$ |
| | | `Isotropic + Tukey corr` | $0.48^{+0.16}_{-0.19}$ | $0.52^{+0.08}_{-0.14}$ | $0.57^{+0.04}_{-0.07}$ | $0.50^{+0.10}_{-0.13}$ |
| 0.5 | 0.62 | `LVK corr` | $0.63^{+0.09}_{-0.08}$ | $0.59^{+0.05}_{-0.05}$ | $0.61^{+0.03}_{-0.02}$ | $0.54^{+0.13}_{-0.04}$ |
| | | `Isotropic + Gaussian corr` | $0.64^{+0.12}_{-0.11}$ | $0.60^{+0.06}_{-0.05}$ | $0.62^{+0.04}_{-0.03}$ | $0.50^{+0.09}_{-0.09}$ |
| | | `Isotropic + Beta corr` | $0.61^{+0.21}_{-0.11}$ | $0.59^{+0.13}_{-0.09}$ | $0.65^{+0.09}_{-0.07}$ | $0.49^{+0.22}_{-0.17}$ |
| | | `Isotropic + Tukey corr` | $0.64^{+0.23}_{-0.25}$ | $0.60^{+0.13}_{-0.08}$ | $0.60^{+0.05}_{-0.03}$ | $0.50^{+0.13}_{-0.15}$ |
| 1 | 0.67 | `LVK corr` | $0.73^{+0.27}_{-0.16}$ | $0.65^{+0.14}_{-0.08}$ | $0.66^{+0.05}_{-0.04}$ | $0.56^{+0.30}_{-0.05}$ |
| | | `Isotropic + Gaussian corr` | $0.62^{+0.23}_{-0.15}$ | $0.59^{+0.11}_{-0.11}$ | $0.61^{+0.06}_{-0.11}$ | $0.49^{+0.09}_{-0.15}$ |
| | | `Isotropic + Beta corr`[(1)] | $0.75^{+1.52}_{-0.47}$ | $0.56^{+0.89}_{-0.21}$ | $0.43^{+0.41}_{-0.11}$ | $0.42^{+1.44}_{-0.19}$ |
| | | `Isotropic + Tukey corr` | $0.63^{+0.45}_{-0.27}$ | $0.58^{+0.16}_{-0.17}$ | $0.59^{+0.05}_{-0.10}$ | $0.50^{+0.08}_{-0.20}$ |

TABLE III. PPD (median and 90% credible interval) of $\cos\tau$ at fixed values of $\cos\tau$ (first column) for the correlated models. The prior and true value are also reported. [(1)] For the `Isotropic + Beta` only, the slice is actually taken at $\cos\tau = 0.99$ numerical issues with the large values that singular beta posteriors can take in the last bin.

| $\cos\tau$ | Truth | Model | 150 events | 500 events | 1500 events | Prior |
|---|---|---|---|---|---|---|
| $\cos\tau \leq -0.98$ | 0.4 | `LVK` | $0.4^{+0.2}_{-0.2}$ | $0.5^{+0.1}_{-0.1}$ | $0.4^{+0.1}_{-0.0}$ | $0.5^{+0.2}_{-0.4}$ |
| | | `Isotropic + Gaussian` | $0.4^{+0.2}_{-0.2}$ | $0.5^{+0.2}_{-0.1}$ | $0.5^{+0.0}_{-0.1}$ | $0.7^{+0.3}_{-0.4}$ |
| | | `Isotropic + Beta` | $0.5^{+0.2}_{-0.3}$ | $0.5^{+0.2}_{-0.2}$ | $0.4^{+0.1}_{-0.3}$ | $0.5^{+6.6}_{-0.4}$ |
| | | `Isotropic + Tukey` | $0.4^{+0.2}_{-0.3}$ | $0.5^{+0.2}_{-0.2}$ | $0.4^{+0.2}_{-0.1}$ | $0.7^{+0.4}_{-0.6}$ |
| $\cos\tau \leq 0.00$ | 38.9 | `LVK` | $36.5^{+6.6}_{-7.9}$ | $41.3^{+3.7}_{-4.2}$ | $40.2^{+2.6}_{-1.2}$ | $42.0^{+7.2}_{-24.3}$ |
| | | `Isotropic + Gaussian` | $37.2^{+7.4}_{-7.7}$ | $41.7^{+4.0}_{-4.5}$ | $40.4^{+3.1}_{-1.6}$ | $50.0^{+15.1}_{-15.8}$ |
| | | `Isotropic + Beta` | $35.0^{+7.6}_{-7.9}$ | $40.7^{+4.3}_{-4.9}$ | $40.4^{+2.3}_{-1.4}$ | $49.9^{+33.1}_{-33.1}$ |
| | | `Isotropic + Tukey` | $37.1^{+8.2}_{-9.5}$ | $41.9^{+4.4}_{-5.1}$ | $41.0^{+1.9}_{-1.5}$ | $50.0^{+26.0}_{-25.3}$ |
| $\cos\tau \geq 0.98$ | 1.0 | `LVK` | $1.1^{+0.4}_{-0.2}$ | $1.0^{+0.2}_{-0.1}$ | $1.0^{+0.2}_{-0.2}$ | $1.0^{+0.9}_{-0.2}$ |
| | | `Isotropic + Gaussian` | $1.0^{+0.3}_{-0.3}$ | $0.9^{+0.2}_{-0.2}$ | $0.9^{+0.1}_{-0.2}$ | $0.7^{+0.3}_{-0.4}$ |
| | | `Isotropic + Beta` | $0.9^{+3.4}_{-0.6}$ | $0.7^{+1.6}_{-0.3}$ | $0.6^{+0.2}_{-0.2}$ | $0.5^{+6.7}_{-0.4}$ |
| | | `Isotropic + Tukey` | $0.9^{+0.6}_{-0.4}$ | $0.9^{+0.2}_{-0.2}$ | $0.9^{+0.1}_{-0.1}$ | $0.7^{+0.3}_{-0.6}$ |

TABLE IV. Median and 90% credible interval on the percentage of sources with $\cos\tau$ in the range given in the first column for the uncorrelated models. Truth and priors are also reported.

| $\cos\tau$ | Truth | Model | 150 events | 500 events | 1500 events | Prior |
|---|---|---|---|---|---|---|
| $\cos\tau \leq -0.98$ | 0.4 | `LVK corr` | $0.5^{+0.2}_{-0.2}$ | $0.6^{+0.1}_{-0.1}$ | $0.5^{+0.1}_{-0.1}$ | $0.6^{+0.2}_{-0.2}$ |
| | | `Isotropic + Gaussian corr` | $0.5^{+0.2}_{-0.2}$ | $0.5^{+0.1}_{-0.2}$ | $0.4^{+0.1}_{-0.1}$ | $0.7^{+0.2}_{-0.2}$ |
| | | `Isotropic + Beta corr` | $0.5^{+0.2}_{-0.2}$ | $0.6^{+0.1}_{-0.1}$ | $0.4^{+0.1}_{-0.1}$ | $0.6^{+3.1}_{-0.3}$ |
| | | `Isotropic + Tukey corr` | $0.5^{+0.2}_{-0.2}$ | $0.6^{+0.1}_{-0.2}$ | $0.4^{+0.1}_{-0.2}$ | $0.7^{+0.2}_{-0.3}$ |
| $\cos\tau \leq 0.00$ | 38.9 | `LVK corr` | $37.5^{+7.9}_{-8.5}$ | $41.4^{+4.5}_{-4.6}$ | $40.4^{+2.2}_{-1.4}$ | $46.2^{+3.5}_{-12.7}$ |
| | | `Isotropic + Gaussian corr` | $38.9^{+9.4}_{-9.2}$ | $41.8^{+4.8}_{-4.6}$ | $39.9^{+1.8}_{-3.1}$ | $50.0^{+8.4}_{-7.4}$ |
| | | `Isotropic + Beta corr` | $36.4^{+7.6}_{-8.1}$ | $40.9^{+4.9}_{-4.7}$ | $40.0^{+2.0}_{-1.0}$ | $50.1^{+15.8}_{-15.4}$ |
| | | `Isotropic + Tukey corr` | $37.1^{+10.9}_{-9.4}$ | $41.3^{+5.2}_{-5.4}$ | $40.8^{+2.1}_{-2.1}$ | $50.0^{+11.5}_{-10.9}$ |
| $\cos\tau \geq 0.98$ | 1.0 | `LVK corr` | $1.1^{+0.4}_{-0.3}$ | $1.0^{+0.2}_{-0.2}$ | $1.0^{+0.1}_{-0.1}$ | $0.9^{+0.5}_{-0.1}$ |
| | | `Isotropic + Gaussian corr` | $0.9^{+0.4}_{-0.3}$ | $0.9^{+0.2}_{-0.2}$ | $0.9^{+0.2}_{-0.1}$ | $0.7^{+0.2}_{-0.2}$ |
| | | `Isotropic + Beta corr` | $1.1^{+3.1}_{-0.7}$ | $0.8^{+1.5}_{-0.3}$ | $0.6^{+0.2}_{-0.1}$ | $0.6^{+3.2}_{-0.3}$ |
| | | `Isotropic + Tukey corr` | $1.0^{+0.7}_{-0.4}$ | $0.9^{+0.2}_{-0.3}$ | $0.9^{+0.1}_{-0.2}$ | $0.7^{+0.2}_{-0.3}$ |

TABLE V. Median and 90% credible interval on the percentage of sources with $\cos\tau$ in the range given in the first column for the correlated models. Truth and priors are also reported.