Improving Posterior Predictive Checks for Gravitational Wave Population Analyses: Interim Report No. 1

Sophia Winney

(Dated: August 2, 2024)

In population analyses of gravitational waves emitted from binary black holes (BBH), spin magnitude and tilt angle distributions provide critical insights regarding BBH evolutionary histories and formation channels. For this reason, developing reliable models of BBH spin component distributions continues to be an essential problem. However, the effects of spin magnitude and directions on gravitational wave signals are subdominant compared to the influence of the effective aligned spin and the effective precessing spin parameters. As a result, obtaining well-constrained measurements of these parameters is difficult, posing challenges when developing models of their astrophysical distributions and determining model accuracy. Posterior predictive checks (PPCs), a widely used model-checking method in gravitational wave science, especially fall short when applied using data with high uncertainties. In this project, we implement alternative data-level PPCs, partial PPCs, and split predictive checks. We aim to explore the efficacy of these methods by applying them to models of varying accuracy of simulated astrophysical populations with the same effective-spin distribution and different spin component distributions. Currently, we have successfully replicated previous results that demonstrate the inability of PPCs to demonstrate model misspecification when the data is highly uncertain.

I. INTRODUCTION

The continued success of gravitational wave observation has permitted population analyses of binary black hole (BBH) mergers [1–3]. Such events can be described by the spin components of the primary and the secondary, entailing each of their spin magnitudes ($\chi_{\{1,2\}}$), azimuthal angles ($\phi_{\{1,2\}}$), and polar angles ($\theta_{\{1,2\}}$) [4]. Previous studies have favored a binary black hole population with small but non-zero spins, as well as a wide range of polar angles. However, these results are highly model dependent; conflicting work has favored a majority of spin zero black holes and remaining nonzero spins that are primarily aligned with orbital angular momentum [5].

These opposing conclusions have significant implications for BBH formation histories. Spin magnitudes, for instance, provide a probe into the angular momentum processes within stellar cores, and large magnitudes can indicate hierarchical black hole formation through previous mergers. Tilts can indicate the formation channel through which a BBH system formed. The isolated binary formation channel consists of binary star systems in which one star becomes a black hole, resulting in mass transfer and an eventual merger in an observable amount of time. These systems tend to exist for long periods of time, conducive to a majority of spins aligned with orbital angular momentum. Conversely, dynamic formation occurs in dense environments where black holes of similar mass congregate and become gravitationally bound into binary systems. Such dynamically formed systems lack tidal effects and mass transfer, thus favoring a random distribution of spin alignment [6]. Hierarchical mergers—which often occur in galactic nuclei, active galactic nuclei, and extremely low-mass ultrafaint dwarf galaxies—may increase BBH eccentricities and quicken inspiral, increasing the likelihood of a merger within Hubble time and further complicating the picture of BBH formation [6].

Although highly significant, the parameters $\chi_{\{1,2\}}$, $\phi_{\{1,2\}}$, and $\theta_{\{1,2\}}$ induce subdominant effects on gravitational wave signals, which are instead primarily influenced by the effective aligned spin (χ_{eff}), containing the spin components that are aligned with the orbital angular momentum, and effective precessing parameter (χ_p), containing the anti-aligned components. The lower dimensionality of informative parameters makes it difficult to ascertain the underlying distributions of spin components for individual gravitational wave events. Because the individual event posteriors are often poorly constrained with high uncertainties, attempting to characterize the hyperposteriors, or the parameters describing the distribution of the individual event parameters (e.g. mean and standard deviation), is difficult. Uncertain data also creates challenges when verifying whether proposed models for these population distributions are in agreement with the observed individual events.

Posterior predictive checks (PPCs) [7], a common test of model accuracy, evaluate the performance of predictive models by checking the consistency between data predicted by the model and current observations. Although widely used in gravitational wave population analyses, PPCs demonstrate significant limitations when looking at uninformative parameters like spin components [8]. The objective of this project is to determine whether alternative types of PPCs (e.g. data-level PPCs, partial predictive checks, and/or split predictive checks) are more discerning tools for model criticism than traditional PPCs. Our approach to posterior predictive checks is described further in Section II, our preliminary results are shown in Sections III and V, and the methods we plan to implement in the coming weeks is detailed in Section IV.

II. METHODS

To ascertain the efficacy of these alternative PPCs, we use the spin component distributions from the simulated astrophysical populations of binary black hole systems from Miller *et al.* [8], which have identical $\chi_{\rm eff}$ distributions but different χ_i and $\cos\theta_i$ distributions. We attempt to recover the known component spin distributions by proposing population models that describe the simulated populations with varying accuracy and then determine whether alternative PPC methods reflect the performance of the models more effectively than regular PPCs.

A. Spin Parameterization

The effective aligned spin [9] is defined as

$$\chi_{\rm eff} = \frac{\chi_1 \cos\theta_1 + q\chi_2\cos\theta_2}{1+q}, \tag{1}$$

and the effective precessing parameter [10] as

$$\chi_{\rm p} = \max\left[\ \chi_1 \sin\theta_1\ ,\ \left(\frac{3+4q}{4+3q}\right) q\chi_2 \sin\theta_2\ \right], \tag{2}$$

where $q = \frac{m_2}{m_1} \leq 1$, χ_1 and χ_2 are the dimensionless component spins, and θ_1 and θ_2 are the angles between the component spins and the orbital angular momentum [11, 12].

B. Simulated Population

The simulated astrophysical distributions consist of three populations, as described in Miller *et al.* [8]. HighSpinPrecessing is characterized by the most extreme spins and tilts, with most spin magnitudes $\chi > 0.5$ and primarily in-plane tilts, indicating significant precession. MediumSpin most closely resembles current spin and tilt constraints. The distribution of spin magnitude peaks and $\chi = 0.25$, with a wide distribution of tilts that are mostly aligned. The majority of LowSpinAligned has spin magnitude $\chi < 0.5$. The distribution is bimodal with peaks at $\cos\theta = 1$ and $\cos\theta = -1$, or perfect alignment and anti-alignment. Individual event parameters are drawn from these distributions, and 70 events with a signal-to-noise ratio greater than 10 are selected to simulate gravitational wave data using IMRPhenomXPHM. The multidimensional posterior for the binary black hole parameters is sampled using Bilby. We then hierarchically model the population of simulated posteriors by using emcee to sample the proposed population model. In this model, the χ distribution is characterized by a beta distribution and the $\cos\theta$ distribution is characterized by a Gaussian.

C. Population Inference Using Posterior Predictive Checks

For each event in our simulated populations, we use Bayesian inference to obtain a posterior distribution on individual-event parameters λ (masses, spins, etc.),

$$P_{\rm PE}(\lambda|d) = \mathcal{L}(d|\lambda)\pi_{\rm PE}(\lambda), \tag{3}$$

with $\pi_{\rm PE}$ indicating the parameter estimation prior. To obtain the posterior according to our proposed model, we reweight $\pi_{\rm PE}$ according to weights $w(\lambda)$ based on the ratio between the proposed population model $\pi_{\rm pop}$ and the parameter estimation priors:

$$\begin{aligned} P_{\rm pop}(\lambda|d) &= \mathcal{L}(d|\lambda)\pi_{\rm pop}(\lambda) \\ P_{\rm pop}(\lambda|d) &= \mathcal{L}(d|\lambda)\pi_{\rm pop}(\lambda)\frac{\pi_{\rm PE}(\lambda)}{\pi_{\rm PE}(\lambda)} \\ P_{\rm pop}(\lambda|d) &= \mathcal{L}(d|\lambda)\pi_{\rm PE}(\lambda)\frac{\pi_{\rm pop}(\lambda)}{\pi_{\rm PE}(\lambda)} \end{aligned}$$

$$P_{\rm pop}(\lambda|d) = P_{\rm PE}(\lambda|d)w(\lambda) \tag{4}$$

where $w(\lambda) = \dfrac{\pi_{\rm pop}(\lambda)}{\pi_{\rm PE}(\lambda)}$. Using these weights, we reweight the individual events according to our proposed priors, creating our "observed" population. We reweight a reference set of detectable injections according to our proposed model, using that population's $P_{\rm draw}$ instead of $\pi_{\rm PE}$, creating our "predicted" population.

For the individual event reweighting, we draw a random set of hyperparameters, calculate weights, and perform a weighted draw to obtain one set of spin components for the individual event. This is repeated 70 times to create an observed catalog. For the injected population, we repeat this process 70 times for each of our 100 predicted catalogs.

If the observed individual-event posteriors have high uncertainties and therefore a widely distributed likelihood, PPCs fall short in identifying inaccurate models. In this scenario, sampling the proposed posterior of a poor model with the draws weighted according to high uncertainties produces nearly identical "observed" and "predicted" populations, making posterior predictive checking unhelpful when dealing with uninformative data. This can be observed in the first and fourth columns of Figure 1 of Section III.

III. CURRENT PROGRESS

A. Population Inference with Inaccurate Models

To become familiar with the data and the process of posterior predictive checks, I recreated Figure 4 of Miller

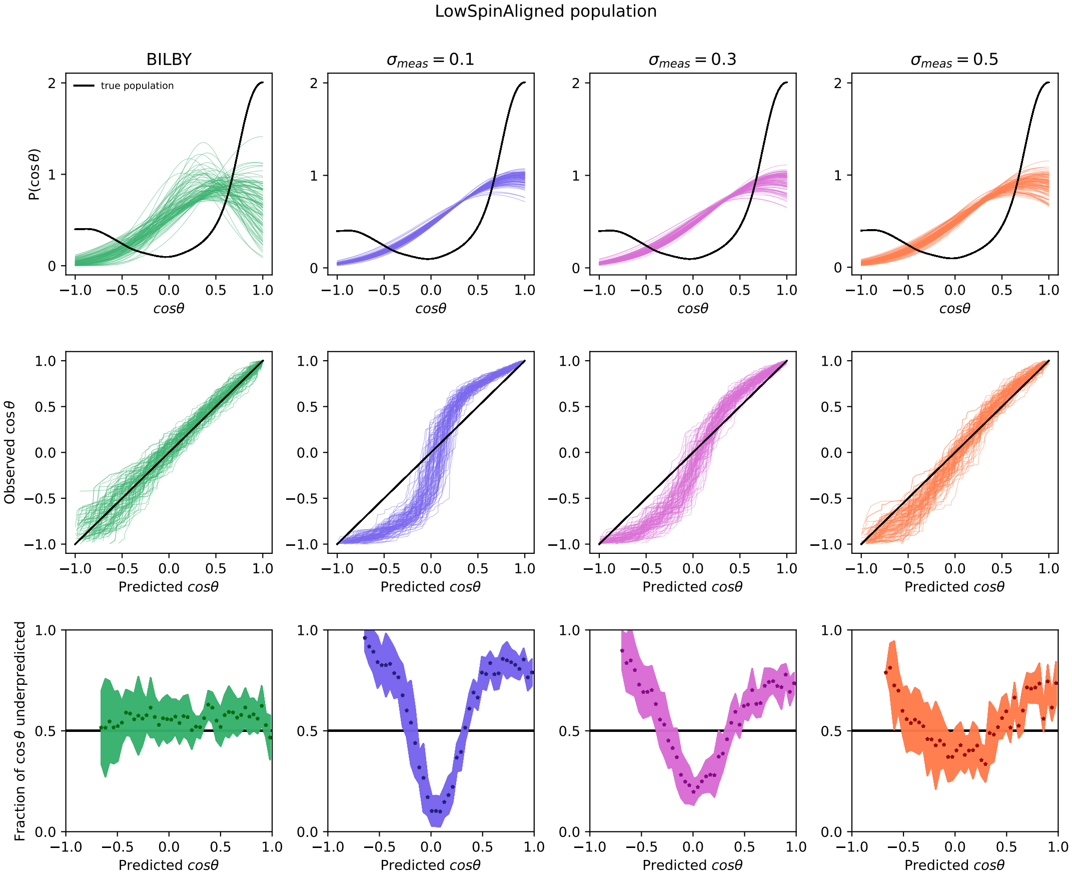

FIG. 1. Results of population inference using posterior predictive draws for $\cos\theta$ of the LowSpinAligned simulated population. **First row**: shows the probability density function for $\cos\theta$ in both the simulated, "true" population and traces obtained from drawing random hyperparameters from the given hyperposterior. **Second row**: alignment between 100 traces of predicted draws and the observed draws from the reweighted individual event population. When uncertainties are high, the inaccuracy of the model becomes difficult to discern. **Third row**: $\cos\theta$ is binned on the x-axis, and points on the y-axis represent the fraction of predicted traces in the corresponding bin that underpredict the true value of $\cos\theta$. The failure of the Gaussian $\cos\theta$ population model to recover the bimodality of the true population is clear with the $\sigma_{\rm meas} = 0.1,\ 0.3$ posteriors but cannot be observed when measurement uncertainties increase.

et al. [8] with the inclusion of the spin magnitude χ and the gravitational wave posterior pipeline Bilby, as shown in Figures 1 and 2. I also created the same visualizations for the MediumSpin and HighSpinPrecessing, presented in Section IV. For each of the three simulated populations, four different posteriors are tested: Bilby, and simulated Gaussian posteriors with uncertainties $\sigma_{\rm meas} = 0.1,\ 0.3,\ 0.5$.

As demonstrated in the top row of Figure 1, all four sets of posteriors are a poor fit for the $\cos\theta$ distribution of the LowSpinAligned population, making this particular simulated population a useful probe into the efficacy of posterior predictive checks.

The second row of Figure 1 shows the alignment between the observed draws and 100 traces from the predicted catalogs. With a low measurement uncertainty of $\sigma_{\rm meas} = 0.1$, the discrepancy between the observed and predicted draws is visually clear. However, when Bilby is used or the measurement uncertainty increases, the individual event posteriors contain more poorly constrained likelihoods, causing the prior to dominate and cause identical observed and predicted populations after reweighting. As a result, the traces more closely follow the diagonal, making it difficult to diagnose the inaccurate model.

Similarly, the third row of Figure 1 shows the fraction of traces that underpredict the probability density values within a binned range of $\cos\theta$. While the $\sigma_{\rm meas} = 0.1,\ 0.3$ posteriors do capture the inability of the model to capture the bimodal nature of the true population distribution, the Bilby and $\sigma_{\rm meas} = 0.5$ posteriors show little scatter around 0.5 and fail to indicate the differences in shape of the true and modeled distributions.

Figure 2 shows the results of posterior predictive checks on the better-constrained parameter χ. Despite the Bilby posterior following the singularly peaked χ distribution more closely than the bimodal $\cos\theta$ distribution, the observed and predicted draws as well as the fraction of traces that underpredict show similar amounts of scatter.

IV. FUTURE OF RESEARCH: ALTERNATIVE POSTERIOR PREDICTIVE CHECKS

The overall objective of this project is to determine whether alternate data-level PPCs, partial PPCs, and/or split predictive checks are more discerning tools for model criticism than typical PPCs. Therefore, in the coming weeks, we plan to implement the following methods on the same simulated populations used in Section III.

A. Data-level Posterior Predictive Checks

The highest level of hierarchical population analysis imposes population level parameters on subsets of individual events and checks for consistency between the subsets to evaluate the accuracy of the population model.

In Section II, we implemented true event-level parameter PPCs, the second level of population alaysis. As discussed in Fishbach *et al.* [13], this includes integrating out the hyperparameters to find a posterior distribution $p(\theta|d_i)$ for the true parameters of a single event based on the data:

$$p(\theta|d_{\mathrm{i}}) = \int p_{\mathrm{pop}}(\theta|\Lambda)p(\Lambda|d_{\mathrm{i}})d\Lambda, \tag{5}$$

where $p(\Lambda|d_{\mathrm{i}})$ is obtained by marginalizing the joint posterior probability distribution of the event level and population level parameters based on the data. This results in an "observed" catalog, and the "predicted" catalog is constructed by weighting the true parameter draws by the detectability of their resulting data. Comparison of these two catalogs also allows for a check of model accuracy.

The lowest level of population analysis is the data-driven PPC. This PPC incorporates measurement uncertainty and detection efficiency to find the predictive posterior probability distribution for a future detection based on the data. The observed parameters are obtained by approximating a single gravitational wave detection with noise and then finding the parameters with the maximum likelihood.

To implement data-level PPCs, we impose the injected hyperparameters while generating an individual event signal rather than reweighting each of the individual events to the proposed model afterward. To do this, we draw random hyperparameters from the hyperposterior and treat these as our "true" parameters. For the Bilby posterior's observed catalog, we first use the true parameters to inject a signal and then recover the spin parameter values with the maximum likelihood according to the generated signal. For the Gaussian posteriors, we draw parameters from a normal distribution centered at the true parameter with the corresponding uncertainty σ_{meas}.

To create the predicted population for Bilby, we take the hyperparameters corresponding to the highest likelihood in the sampler output. For the Gaussian posteriors, we create a kernel density estimator for the parameters and select the value of maximum likelihood.

Ideally, this method would skirt the issues of traditional PPCs by replacing the step of reweighting uncertain individual events with the injection of drawn hyperparameters.

B. Partial Posterior Predictive Checks

Training the improper prior into a distribution that integrates to 1 and evaluating measures of surprise with the same data can lead to a non-representative p-value. Partial PPCs address this shortcoming of posterior predictive checking by avoiding the double-use of data. The partial PPC method does use the statistic data t_{obs} to compute measures of surprise, but only uses information not present in t_{obs} when training the prior. The conditional distribution $f(x_{\mathrm{obs}}|t_{\mathrm{obs}}, \mathbf{P})$ is used as the likelihood to determine the posterior distribution π_{ppp} of parameters $\mathbf{P}$,

$$\begin{aligned}\pi_{\mathrm{ppp}}(\mathbf{P}|x_{\mathrm{obs}}\backslash t_{\mathrm{obs}}) &\propto f(x_{\mathrm{obs}}|t_{\mathrm{obs}}, \mathbf{P})\pi(\mathbf{P}) & (6)\\ &\propto \frac{f(x_{\mathrm{obs}}|\mathbf{P})\pi(\mathbf{P})}{f(t_{\mathrm{obs}}|\mathbf{P})}, & (7)\end{aligned}$$

which is then used as a prior to determine posterior of t. With the contribution of t_{obs} already eliminated, $\mathbf{P}$ is integrated out of the posterior:

$$m_{\mathrm{ppp}}(t|x_{\mathrm{obs}}\backslash t_{\mathrm{obs}}) = \int f(t|\mathbf{P})\pi(\mathbf{P}|x_{\mathrm{obs}}\backslash t_{\mathrm{obs}})d\mathbf{P} \tag{8}$$

The new p-value then takes the form

$$p = \mathrm{Pr}^{m_{\mathrm{ppp}}(t|x_{\mathrm{obs}}\backslash t_{\mathrm{obs}})}(t(\mathbf{x}) \geq t(\mathbf{x}_{\mathrm{obs}})). \tag{9}$$

The use of $f(x_{\mathrm{obs}}|t_{\mathrm{obs}}, \mathbf{P})$ to define the likelihood instead of $f(x_{\mathrm{obs}}|\mathbf{P})$ evades the double-use of data that occurs in PPCs when training the prior into a proper distribution and then evaluating the p-value [14, 15].

C. Split Predictive Checks

The split predictive check (SPC) similarly aims to avoid repeated use of data. Rather than conditioning the prior for $\mathbf{P}$ on the influence of t_{obs} (as in partial PPCs), however, split predictive checks partition the data into two disjoint subsets from the start. With a single split of data $x_{\mathrm{obs}} = x_a + x_b$, the method uses different subsets

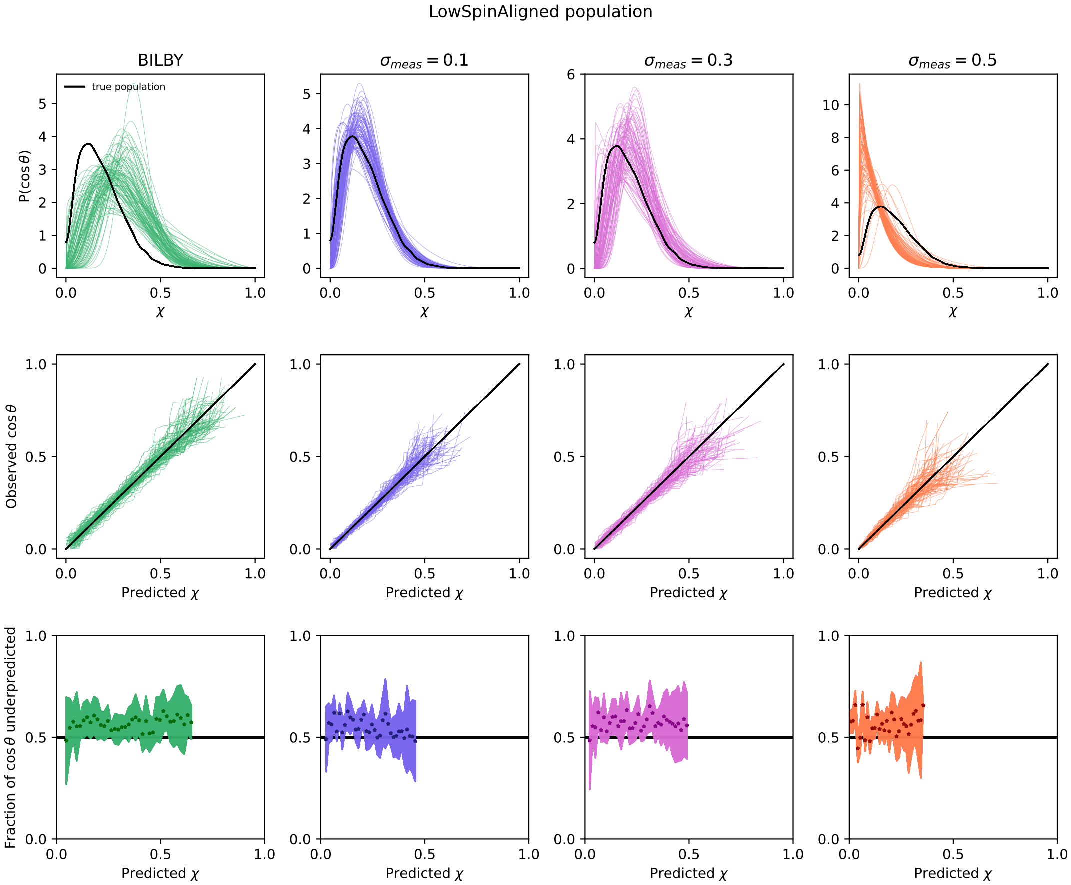

FIG. 2. Results of population inference using posterior predictive draws for χ of the LowSpinAligned simulated population. χ is better constrained within the true population, causing indicators of model misspecification to be more present in the second and third rows.

when training the posterior and when determining the p-value. The posterior distribution for x under the null model

$$m_{\text{SPC}}(x_{\text{pred}}|x_a) = \int f(x_{\text{pred}}|\mathbf{P})\pi(\mathbf{P}|x_a)d\mathbf{P} \tag{10}$$

integrates out the parameters and is used to define a new p-value

$$p = \text{Pr}^{m_{\text{SPC}}(x_{\text{pred}}|x_a)}(t(x_b) \geq t(x_{\text{pred}})). \tag{11}$$

The divided split predictive check (divided SPC) extends this method. Data is divided into N equal subsets, and the single SPC p-value is calculated for each individual subset. The divided SPC p-value is defined as the p-value obtained by performing the Kolmogorov–Smirnov test for uniformity on the collection of single SPC p-values [16].

V. APPENDIX A: RESULTS OF POSTERIOR PREDICTIVE CHECKS FOR THE HIGHSPINPRECESSING AND MEDIUMSPIN SIMULATED POPULATIONS

We used the same approach described in Section II on the HighSpinPrecessing and the MediumSpin populations. The results are shown below in Figures 3 and 4.

VI. ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation Research Experience for Undergraduates (NSF REU) program, the LIGO Laboratory Summer Undergraduate Research Fellowship program (NSF LIGO), and the California Institute of Technology Student-Faculty Programs. Thank you to Simona Miller and Dr. Katerina Chatziioannou for their support and guidance.

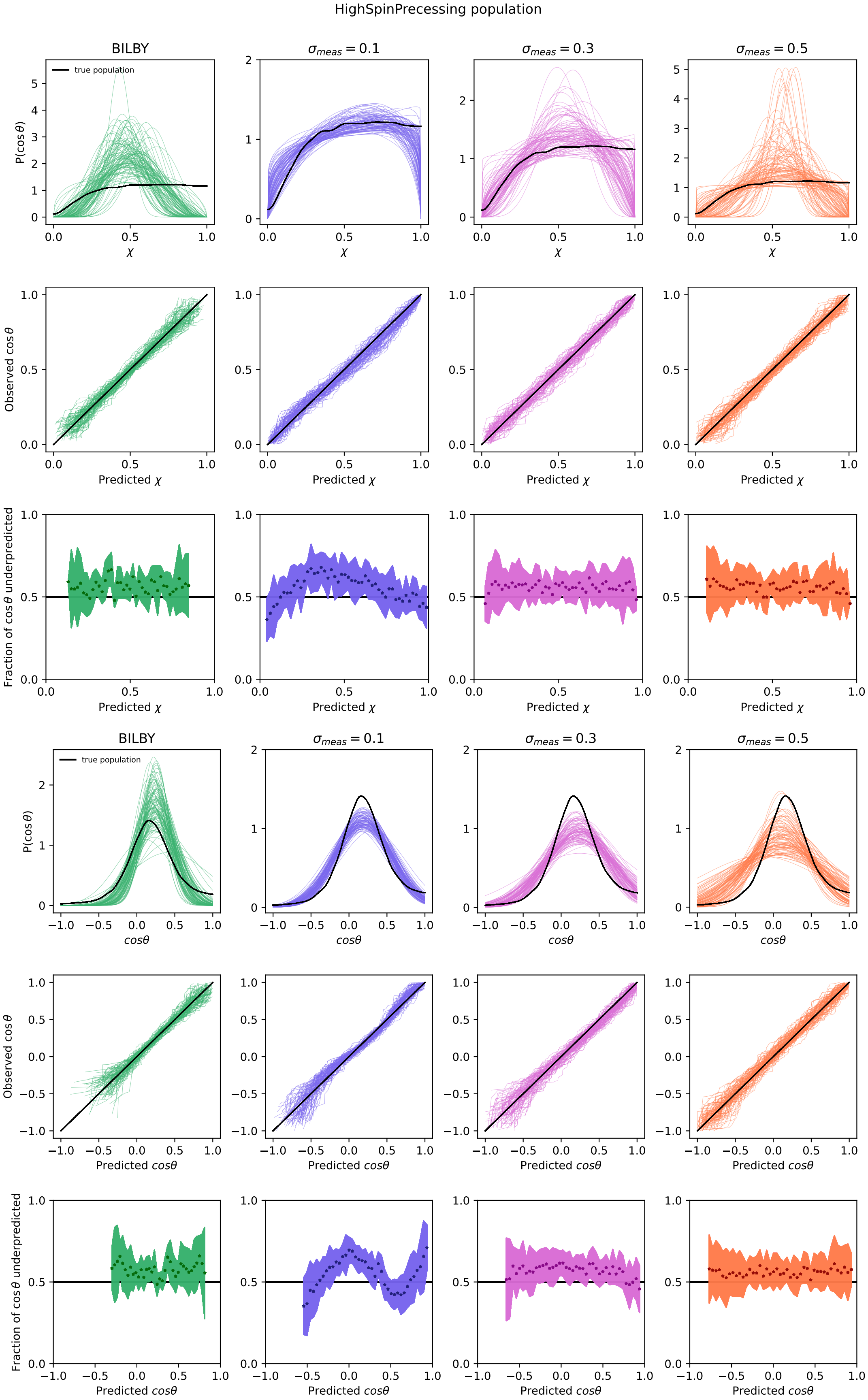

FIG. 3. Results of population inference using posterior predictive draws for χ and $\cos\theta$ of the HighSpinPrecessing simulated population.

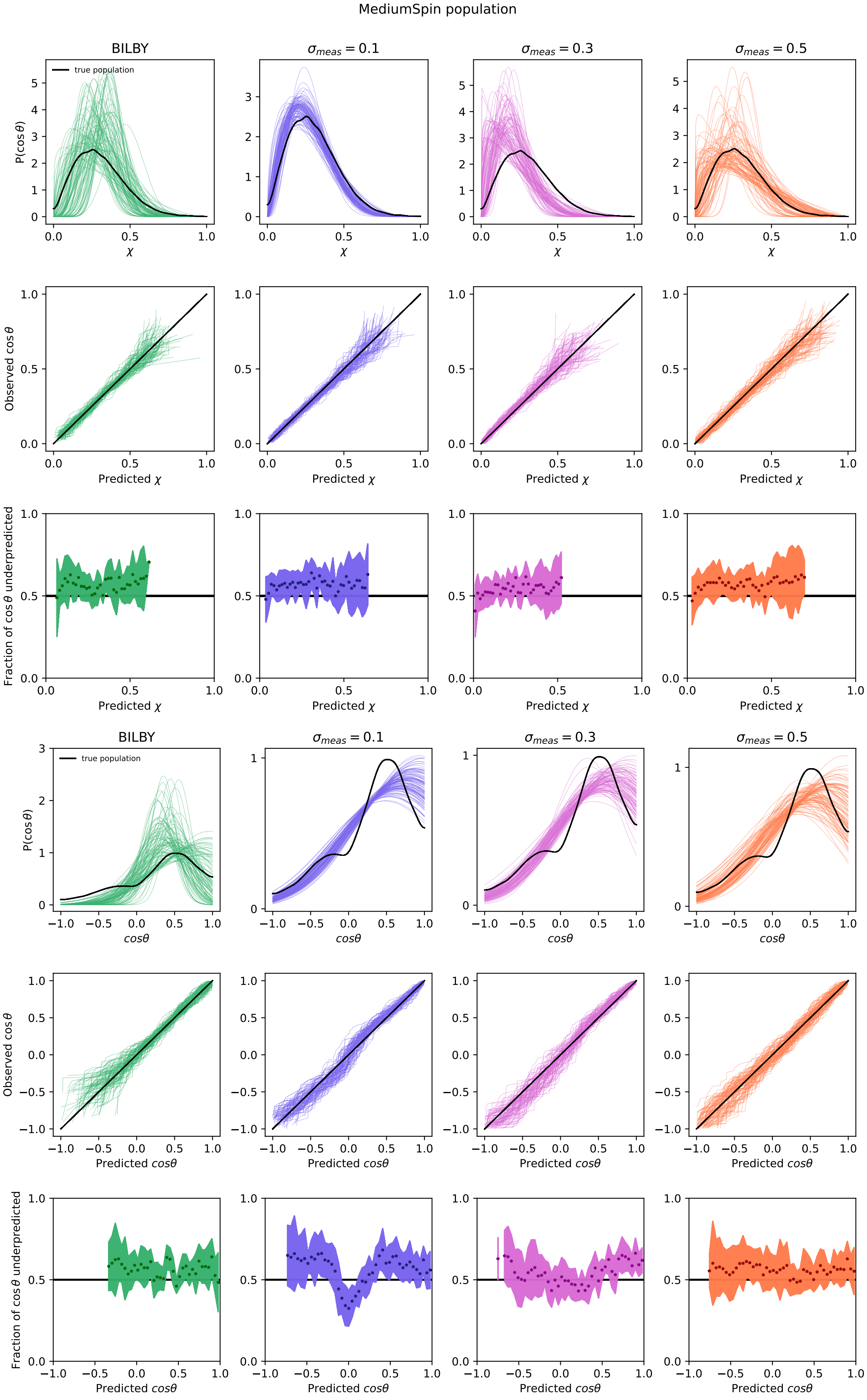

FIG. 4. Results of population inference using posterior predictive draws for χ and $\cos\theta$ of the MediumSpin simulated population.

[1] R. Abbott, **882**, L24 (2019).

[2] R. Abbott, The Astrophysical Journal Letters **913**, 10.3847/2041-8213/abe949 ().

[3] R. Abbott, The population of merging compact binaries inferred using gravitational waves through gwtc-3 ().

[4] J. Roulet and T. Venumadhav, Inferring binary properties from gravitational wave signals (2024), arXiv:2402.11439 [stat.ME].

[5] T. A. Callister, S. J. Miller, K. Chatziioannou, and W. M. Farr, The Astrophysical Journal Letters **937**, L13 (2022).

[6] 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, The Astrophysical Journal **910**, 152 (2021).

[7] M. J. Bayarri and M. E. Castellanos, Statistical Science **22**, 322 (2007).

[8] S. J. Miller, Z. Ko, T. A. Callister, and K. Chatziioannou, Gravitational waves carry information beyond effective spin parameters but it is hard to extract (2024), arXiv:2401.05613 [gr-qc].

[9] Racine, Physical Review D **78**, 10.1103/physrevd.78.044021 (2008).

[10] P. Schmidt, F. Ohme, and M. Hannam, Physical Review D **91**, 10.1103/physrevd.91.024043 (2015).

[11] D. Gerosa, M. Mould, D. Gangardt, P. Schmidt, G. Pratten, and L. M. Thomas, Physical Review D **103**, 10.1103/physrevd.103.064067 (2021).

[12] L. M. Thomas, P. Schmidt, and G. Pratten, Physical Review D **103**, 10.1103/physrevd.103.083022 (2021).

[13] M. Fishbach, W. M. Farr, and D. E. Holz, The Astrophysical Journal Letters **891**, L31 (2020).

[14] M. J. Bayarri and J. O. Berger, Journal of the American Statistical Association **95**, 1127 (2000).

[15] J. M. Robins, A. van der Vaart, and V. Ventura, Journal of the American Statistical Association **95**, 1143 (2000).

[16] F. J. Massey Jr., Journal of the American Statistical Association **46**, 68 (1951).