Improving Posterior Predictive Checks for Gravitational Wave Population Analyses: Interim Report No. 2

Sophia Winney

(Dated: August 14, 2024)

In population analyses of gravitational waves emitted by merging binary black holes (BBH), spin magnitude and tilt angle distributions provide key insights regarding BBH evolutionary histories and formation channels. Therefore, developing reliable BBH spin population models is essential. However, the effects of spin magnitude and tilt on gravitational wave signals are subdominant. Measurements of these parameters tend to be poorly constrained, posing challenges when assessing the accuracy of proposed population models. Posterior predictive checks (PPCs), a widely used model-checking method in gravitational wave science that compares observed data to the population model's predictions, demonstrate limitations when used on events with high uncertainties. We implement data-level PPCs and partial PPCs on simulated populations with known underlying parameter distributions to determine whether they perform better than traditional PPCs when evaluating inaccurate models. We have demonstrated the inability of traditional PPCs to identify inaccurate models when individual event measurements are highly uncertain. Additionally, we have shown that data-level PPCs are more discerning of deviations of the observed data from inaccurate model predictions than traditional PPCs. With certain choices of test statistics, data-level PPC p-values better reflect model misspecification, and traditional PPC p-value distributions can help to infer characteristics of the true population..

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 (more massive) and the secondary (less massive) black hole (BH), entailing each of their spin magnitudes ($\chi_{\{1,2\}}$), azimuthal angles ($\phi_{\{1,2\}}$), and polar (i.e. "tilt") angles ($\theta_{\{1,2\}}$) [4]. Previous studies have found that the BBH population has preferentially small but non-zero spins, as well as a wide range of tilt angles [5]. 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 [6] [7].

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. If angular momentum transport within stars is efficient, then spin magnitudes at the formation of BHs are theorized to be small [8]. BBH systems formed through hierarchical mergers (as opposed to stellar collapse) have larger spin magnitudes because the spin magnitudes of BHs from previous mergers add constructively when forming the remnant black hole [9].

Spin tilts provide further indication of 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 [10].

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, attempting to characterize the posteriors, or the parameters describing the distribution of the individual event parameters across the population (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, i.e. whether or not a selected population model is a good fit to the observed data.

Posterior predictive checks (PPCs) [11], 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, traditional PPCs demonstrate significant limitations when looking at uninformative parameters like spin components [12]. 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 is described further in Section II, our preliminary results are shown in

Sections III and Appendix A, and the methods we plan to implement in the coming weeks are detailed in Section IV.

II. METHODS

To ascertain the efficacy of alternative types of PPCs, we use the spin component distributions from the simulated astrophysical populations of binary black hole systems from Miller *et al.* [12], which have identical $\chi_{\rm eff}$ distributions but different χ_i and $\cos\theta_i$ distributions. We attempt to recover the known injected 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 traditional PPCs.

In the remainder of the Methods section, we present the process through which we obtain our simulated population and the different types of PPCs we employ.

A. Simulated Population

The simulated astrophysical distributions consist of three populations, as described in Miller *et al.* [12]. 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. These three populations are not astrophysically motivated, but rather are selected because they have qualitatively different component spins that all produce the same $\chi_{\rm eff}$ distribution. The individual event posteriors were constructed using simulated Gaussian measurement uncertainties with $\sigma_{\rm meas} = 0.1, 0.3, 0.5$ or the parameter estimation software Bilby, which averages a measurement uncertainty $\sigma_{\rm meas} \approx 0.48$ across all runs [13].

For each population, individual event parameters are drawn from the respective underlying distribution, and 70 events with a network signal-to-noise ratio greater than 10 are selected to simulate gravitational wave data using the phenomenoligical waveform model IMRPhenomXPHM [14]. The multidimensional posterior for the binary black hole parameters is sampled using Bilby [13]. We then hierarchically model the population of simulated posteriors by using the Python package `emcee` to sample the hyperparameters of the proposed population model [15]. Here, the χ distribution is characterized by a Beta distribution and the $\cos\theta$ distribution is characterized by a truncated Gaussian. The Gaussian population model is intended fit the simulated $\cos\theta$ distributions poorly, especially the LowSpinAligned population, which will allow us to evaluate the performance of PPCs when characterizing a bad population model. The hyperparameters sampled were the mean and standard deviation of each parameter.

B. Population Inference Using Posterior Predictive Checks

There are three levels of hierarchical population analysis on which PPCs can be performed. 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. [16]

In Section II B 1, we implement traditional parameter PPCs, the second level of population analysis and what is traditionally done in LIGO data analysis. As discussed in Fishbach *et al.* [16], this includes integrating out the hyperparameters Λ to find a posterior distribution $p(\theta|x_i)$ for the true parameters θ of a single event based on the data x:

$$p(\theta|x) = \int \mathcal{L}(\theta|\Lambda)\pi(\Lambda|x)d\Lambda, \tag{1}$$

where $p(\Lambda|d_{\rm i})$ 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. The “predicted” catalog against which we compare this observed catalog is constructed by weighting a set of true parameter draws by the detectability of their resulting data according to the same threshold for detection as used on the real data analysis pipeline that produces the observed catalogs. Comparison of these two catalogs also allows for a check of model accuracy.

The third and lowest level of a population analysis is the data itself. Model checking on this level corresponds to a data-driven PPC, conducted here on the maximum likelihood (rather than true) values of indiviudal-event parameters θ, which we implement in Section II B 2. This PPC incorporates measurement uncertainty and detection efficiency to find the predictive posterior probability distribution for a future detection based on the data.

1. *Traditional Posterior Predictive Checks*

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

$$p_{\rm PE}(\theta|x) = \mathcal{L}(x|\theta)\pi_{\rm PE}(\theta), \tag{2}$$

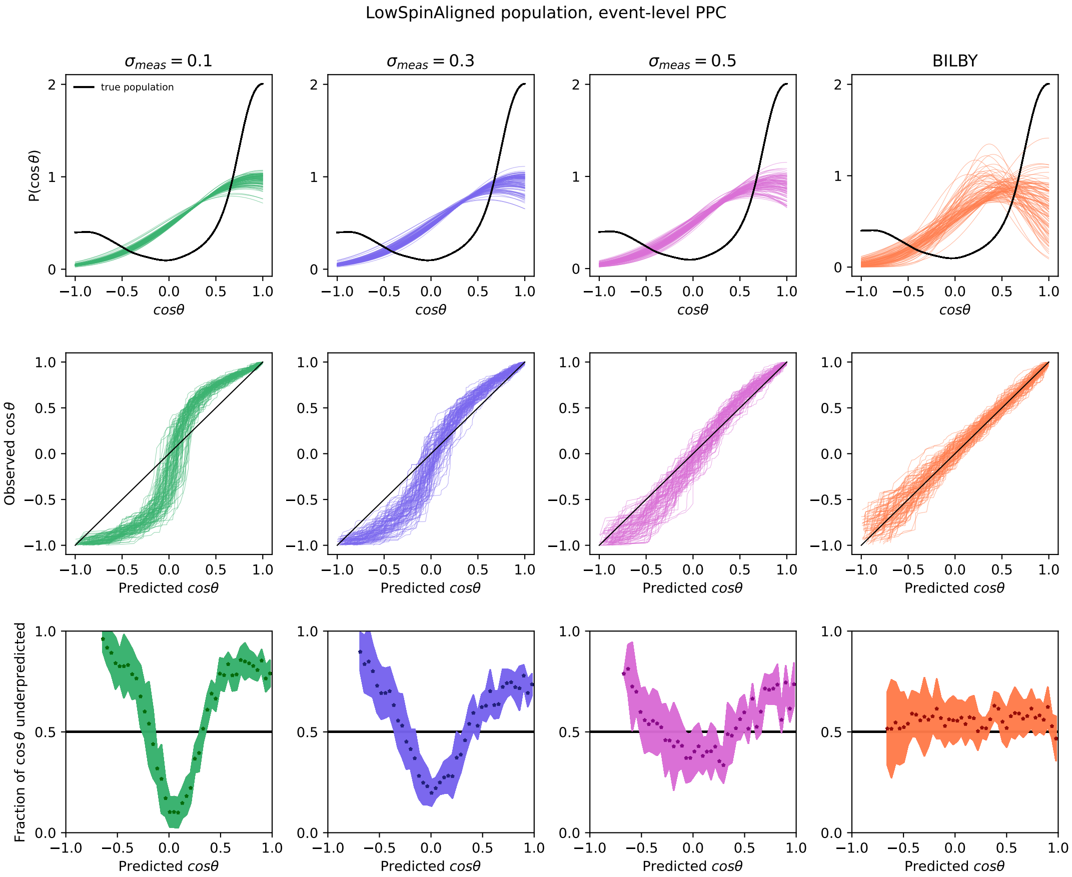

FIG. 1. Results of population inference using posterior predictive draws for $\cos\theta$ of the LowSpinAligned simulated population from Miller *et al.* [12]. Each column shows the results for the different sets of individual event posteriors corresponding to Gaussians with measurement uncertainty $\sigma_{\rm meas}$ or Bilby. [12] **First row**: The probability density function for $\cos\theta$ in both the simulated, "true" population (black) and traces obtained from drawing random hyperparameters from the given posterior (colors). **Second row**: 100 traces of predicted draws plotted against 100 of the observed draws from the reweighted individual event population. The more closely the traces follow the diagonal, the more indicative that the observed and predicted catalogs were drawn from the same distribution, with a diagonal trace implying a perfect fit. 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 $\cos\theta$. To calculate this for each bin, we find the fraction of traces in the second row with a slope less than one. 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 uncertainty increases.

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 $w(\theta)$ based on the ratio between the proposed population model $\pi_{\rm pop}$ and the parameter estimation priors:

$$p_{\rm pop}(\theta|x) = \mathcal{L}(x|\theta)\pi_{\rm pop}(\theta)$$

$$p_{\rm pop}(\theta|x) = \mathcal{L}(x|\theta)\pi_{\rm pop}(\theta)\frac{\pi_{\rm PE}(\theta)}{\pi_{\rm PE}(\theta)}$$

$$p_{\rm pop}(\theta|x) = \mathcal{L}(x|\theta)\pi_{\rm PE}(\theta)\frac{\pi_{\rm pop}(\theta)}{\pi_{\rm PE}(\theta)}$$

$$p_{\rm pop}(\theta|x) = p_{\rm PE}(\theta|x)w(\theta) \tag{3}$$

where $w(\theta) = \dfrac{\pi_{\rm pop}(\theta)}{\pi_{\rm PE}(\theta)}$. 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.

We perform the following steps to reweight our individual events and build an observed catalog:

1. Draw a random set of hyperparameters.
2. Calculate the probability for each parameter value across the individual event posterior using the drawn hyperparameter and the population model.
3. Calculate weights using the ratio of the population

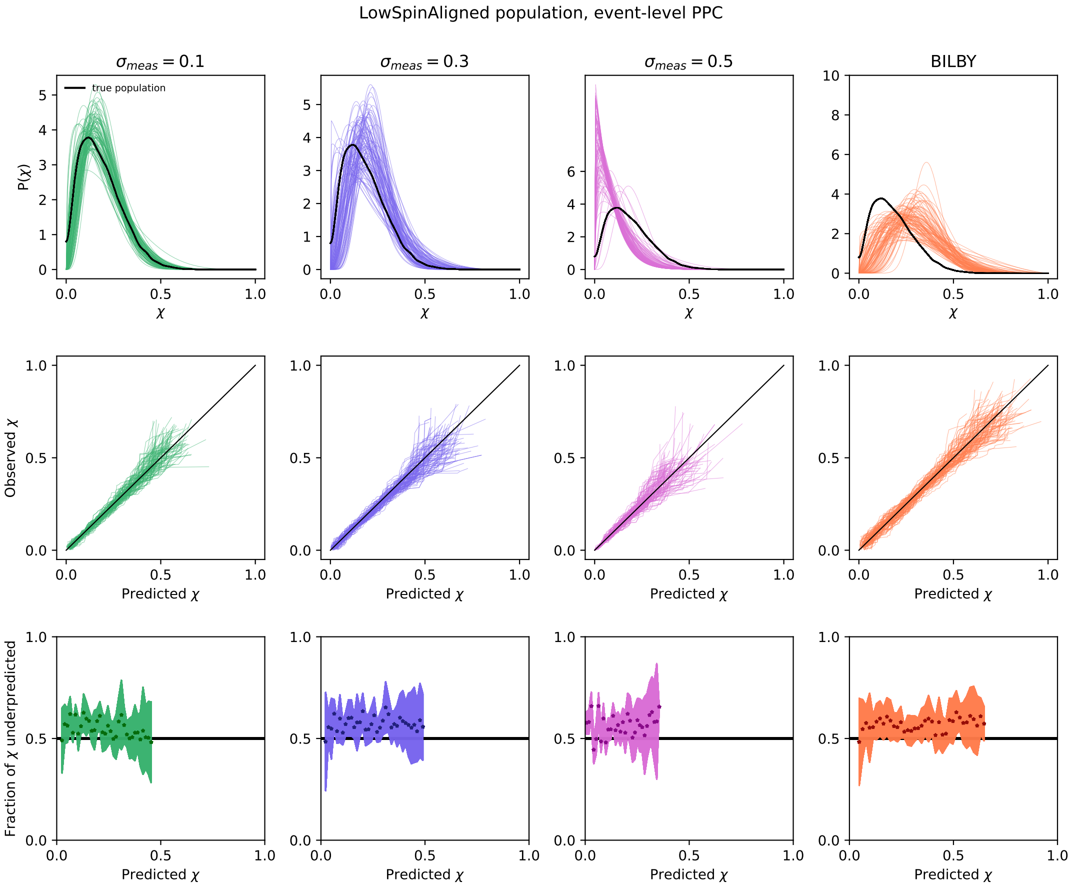

FIG. 2. Same as Fig. 1 but for the spin magnitude χ. χ is better constrained than $\cos\theta$ within the true population, causing indicators of model misspecification to be more present in the second and third rows.

model prior and the individual event parameter estimation prior.

4. Perform one weighted draw from the individual event posterior.

5. Repeat 70 times, once for each individual event.

To create a predicted catalog, we:

1. Draw a random set of hyperparameters.

2. Calculate the probability for each parameter value across the injected population using the drawn hyperparameter and the population model.

3. Calculate weights using the ratio of the population model prior and the injected population parameter estimation prior.

4. Perform one weighted draw from the injected population.

5. Repeat 70 times.

We repeat the above steps to obtain 1000 realizations of predicted and observed 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, for example, in the third and fourth columns of Fig. 1.

2. *Data-Level Posterior Predictive Checks*

Ideally, data-level PPCs would skirt the issues of traditional PPCs by avoiding the step of reweighting uncertain individual events. Rather than reweighting each individual event and drawing a random value from the reweighted posterior to be the true θ, we select the value of maximum likelihood of θ for each event in the observed catalog. For the predicted catalogs, we draw a set of parameters from our population model (using the reweighted injected population) and use them to emulate the detection of a gravitational wave signal in Gaussian

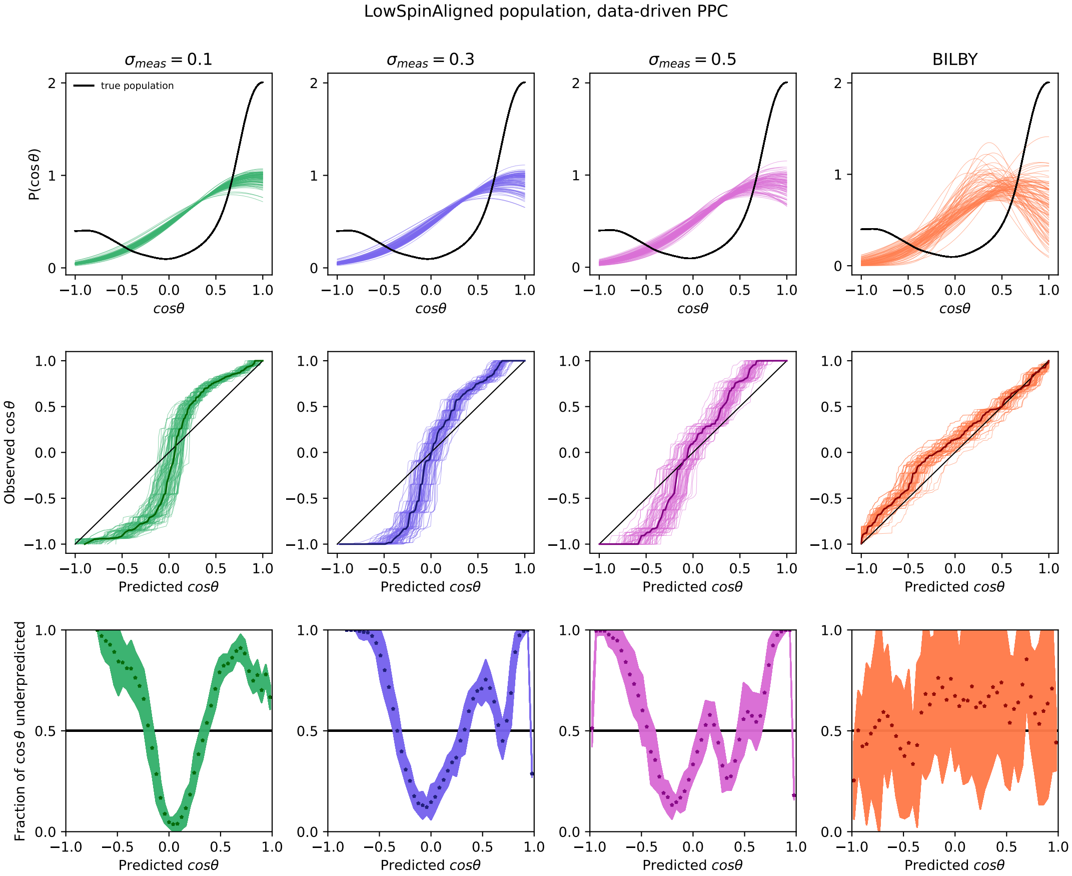

FIG. 3. Results of population inference using data-driven posterior predictive draws for $\cos\theta$ of the LowSpinAligned simulated population. Each column shows the results for the different posteriors corresponding to Gaussians with measurement uncertainty $\sigma_{\rm meas}$ or Bilby. **First row**: The probability density function for $\cos\theta$ in both the simulated, "true" population and traces obtained from drawing random hyperparameters from the given posterior. **Second row**: 100 traces of predicted draws plotted against 100 of the observed draws from the reweighted individual event population, where proximity to the diagonal implies the population model better fit. All observed draws are the maximum likelihood parameter of an individual event posterior, meaning the observed catalog for each trace is the same. The darker trace contains the average of the predicted draws corresponding to each observed draw. In comparison to the traditional PPCs, the traces are more similar to each other, making deviations from the diagonal more apparent. **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 $\cos\theta$. To calculate this for each bin, we find the fraction of traces in the second row with a slope less than one. With each posterior, model misspecification is visually apparent. Unlike in Fig. 1, there are changes in the slope and concavity of the curve that show a high sensitivity to the local behavior of the traces. The wide 3σ range for the Bilby posterior is a result of having used only 100 predicted catalogs rather than 1000, as with the other posteriors.

noise. From each of these signals, we find the maximum likelihood θ. For the Gaussian posteriors, this is done by centering a Gaussian distribution with $\sigma_{\rm meas}$ on the drawn parameters and then drawing one value. For the Bilby posterior, we use Bilby to inject a signal and then find the maximum likelihood parameters using the `FisherMatrixPosteriorEstimator` class [17].

III. CURRENT PROGRESS

A. Results of the Traditional Posterior Predictive Check

We recreated Figure 4 of Miller *et al.* [12] with the inclusion of the spin magnitude χ and the gravitational wave posterior pipeline Bilby, as shown in Figures 1 and 2. We also created the same visualizations for the MediumSpin and HighSpinPrecessing, presented in Appendix A; see Figs. 10 and 11. For each of the three simulated populations, four different posteriors are tested: Bilby, and simulated Gaussian posteriors with measure-

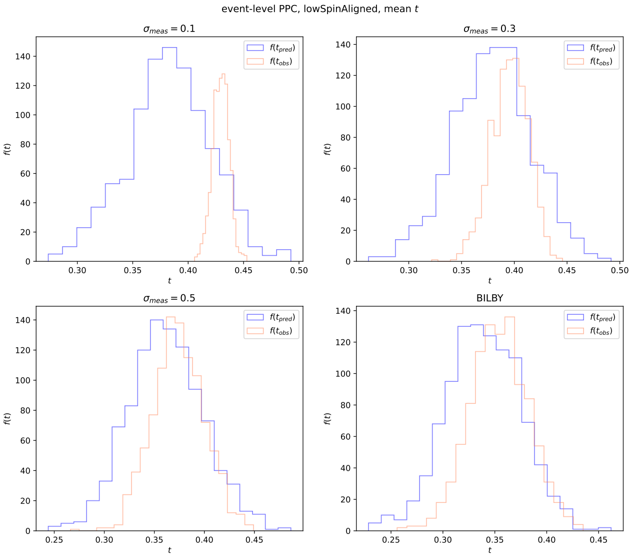

FIG. 4. The observed and predicted distributions of the test statistic t when defined as the mean for the 1000 observed (shown in pink) and predicted (shown in blue) catalogs from the traditional PPC. As measurement uncertainties increase, $f(t_{\rm obs})$ becomes more similar to $f(t_{\rm pred})$ due to the reweighting of uncertain individual event posteriors.

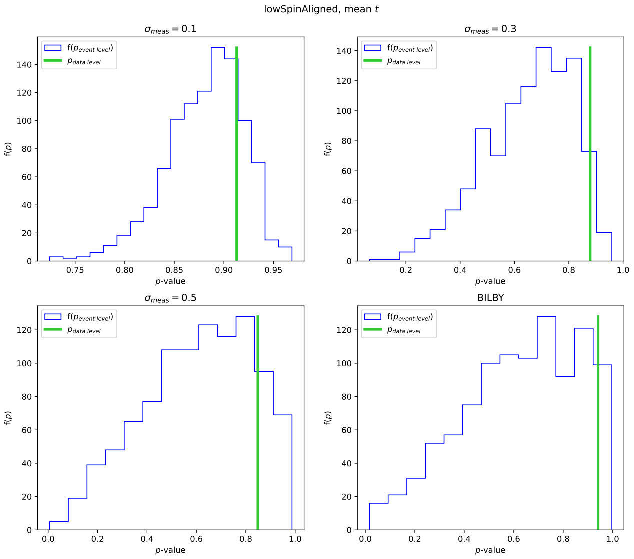

FIG. 5. The distribution of p-values for each of the traditional observed catalogs is shown by the blue histogram. The data-level PPC process produces observed catalogs that are identical, resulting in only one p-value, shown by the green line. The distribution of p-values becomes more uniform on $(0, 1)$ as the measurement uncertainty increases; uniformity of p-values would misleadingly suggest that the observed samples are drawn from the null distribution. The data-level p-value always falls at $\approx 0.85 - 0.95$, indicating that the observed catalog is not compatible with the model, which tends to underpredict t.

ment uncertainties $\sigma_{\rm meas} = 0.1,\ 0.3,\ 0.5$, corresponding to the standard deviations of individual event posteriors. The parameters of actual gravitational wave detections constrained by BILBY parameter estimation have, on average, a measurement certainty $\sigma_{\rm meas} > 0.5$

As demonstrated in the top row of Fig. 1, when assuming a Gaussian population model, all four sets of posteriors yield a poor fit for the bimodal $\cos\theta$ distribution of the LOWSPINALIGNED population, making this particular simulated population a useful probe into the efficacy of posterior predictive checks.

The traces in the second row of Fig. 1 show 100 observed and 100 predicted catalogs sorted and plotted against each other. Samples taken from the same distribution will on average follow a 1:1 ratio, meaning traces closer to the diagonal indicate a good fit between the model and the data. With a low individual-event 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, analogously, 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 prior reweighting. As a result, the traces more closely follow the diagonal, making it difficult to diagnose the inaccurate model.

Similarly, the third row of Fig. 1 shows the fraction of traces that underpredict the probability density values within a binned range of $\cos\theta$. For each bin, the underprediction is obtained by calculating the fraction on traces with a slope below 1. 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.

Fig. 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.

B. Results of the Data-Level Posterior Predictive Check

For comparison to the traditional posterior predictive check, we created the same visualizations as described in Section III A for the results of the data-level PPC on the LOWSPINALIGNED $\cos\theta$ population, which is shown in Fig. 3. We use the same simulated population and population model; as a result, the top row of the figure is the same as in Fig. 1.

The middle row of Fig. 3 shows 100 predicted and 100 observed catalogs sorted and plotted against each other, where traces closer to the diagonal implying a better fit.

However, because each observed event is drawn from the maximum likelihood value of the parameter distribution, each observed catalog contains the same set of parameters. We also calculated an average curve by taking the mean of the predicted values that correspond to each observed event. The traces for the data-level PPC follow each other more closely than in the traditional PPC, making deviations from the 1:1 ratio more visually apparent. For the σ_{meas} posterior especially, the data-level PPC shows significantly more non-diagonality, properly indicating the poor fit between the model and the data. Therefore, data-level PPCs are a better tool than traditional PPCs for qualitatively assessing population model inaccuracies. Deviations from the diagonal are also more apparent in the Bilby case, but it is unclear whether this is due to better performance from the data-level PPC or the inability of `FisherMatrixPosteriorEstimator` class to recover the injected parameters.

The bottom row of Fig. 3 shows the fraction of predicted catalogs in bins of $\cos\theta$ values that underpredict the observed value. The data-level PPCs show more discrepancy between the population model and the simulated population (especially for the $sigma_{\text{meas}} = 0.5$ case). The data-level PPC also exhibits a much higher sensitivity to local features because the observed events are fixed, causing less variance in the slope. In the $\cos\theta \approx 0 - 0.5$ range, the curve shows oscillations around the $y = 0.5$ line that do not correspond to any characteristics of the population model or the simulated population shown in the top row. While the peaks and lows of the bottom row in Fig. 1 would imply a bimodality in the true population that the population model misses, such a statement could not be inferred from Fig. 3. Therefore, the data-level PPCs are both better at indicating bad models and highly sensitive to local fluctuations in the traces.

The wide range for the 3σ confidence interval for the Bilby posterior on the bottom row of 3 is a result of only 100 predicted catalogs being created rather than 1000, as with the other posteriors. This is due to the long runtime of Bilby's FisherMatrixPosteriorEstimator; we will run this maximum likelihood finder to further constrain the fraction underpredicted curve by the end of the program.

C. p-values as a Probe for Model Misspecification

The p-value, a useful tool for assessing the goodness of fit of models, reflects the probability that a sample could be observed assuming the proposed population model is true. In our case, determining the p-value for an observed catalog requires the choice of a test statistic t, which ideally reflects a significant characteristic of the population we are trying to model. We here use four choices of t: the distribution's mean, the distribution's standard deviation, the ratio of aligned spins to anti-aligned spins, and the ratio of aligned spins to zero spins.

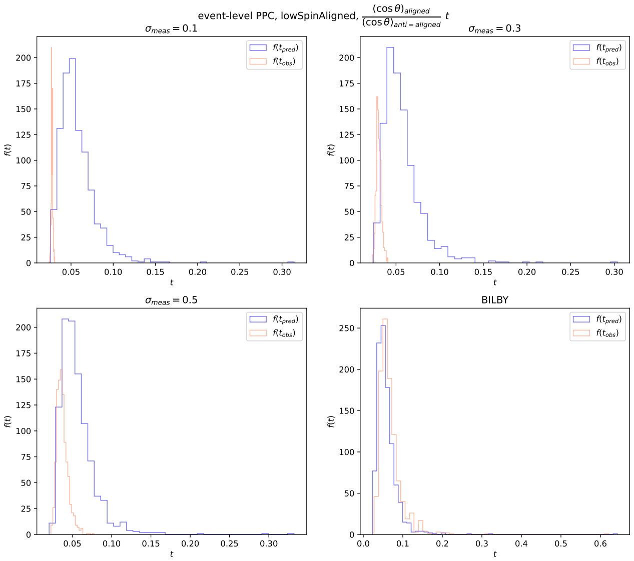

FIG. 6. The observed and predicted distributions of the test statistic t when defined as the number of aligned spins ($0.33 < \cos\theta \leq 1$) divided by the number of anti-aligned spins ($-1 \leq \cos\theta < -0.33$) from the traditional PPC. The $f(t_{\text{obs}}$ distribution is shown in pink and the $f(t_{\text{pred}})$ distribution is shown in blue. The distribution of t_{obs} falls primarily on the left tail of the $f(t_{\text{pred}})$ distribution for the Gaussian posteriors. For the Bilby posteriors, the two distributions are closely aligned due to the reweighting of individual events with Bilby's higher uncertainties.

To calculate the p-value for an observed catalog, we first calculate t_{pred} for each of the 1000 predicted catalogs and used a kernel density estimator to find the distribution $f(t_{\text{pred}})$ across all of the catalogs. We then calculate t_{obs} for the catalog and integrate $f(t_{\text{pred}})$ on the interval $[-\infty, t_{\text{obs}}]$. Therefore, a higher deviation of the p-value from $p = 0.5$ indicates more incompatibility between the model and the observed data. [11]

We first use the catalog's ***mean*** as a test statistic. The mean could reflect whether spins tend to be aligned, as in the case of isolated binary formation, or uniformly distributed, as in the case of dynamic formation. However, in the case of a bimodal $\cos\theta$ population (such as the LowSpinAligned simulated population), this statistic could yield misleading results. The distributions of the means for the traditional, traditional PPCs are shown in 4. When $\sigma_{\text{meas}} = 0.1$, it is clear that the means for the observed catalogs lie on the right of the distribution $f(t_{\text{pred}})$, and the tendency of the incompatible model to underpredict the mean is apparent. With higher σ_{meas}, the distributions of t become less distinguished.

A comparison of the p-values from the means of the traditional and the data-level PPCs is shown in Fig. 5. The blue histogram represents the distribution of p-values across all observed catalogs for the traditional PPC. As measurement uncertainties increase, the p-values become more uniformly distributed, implying a good model. The

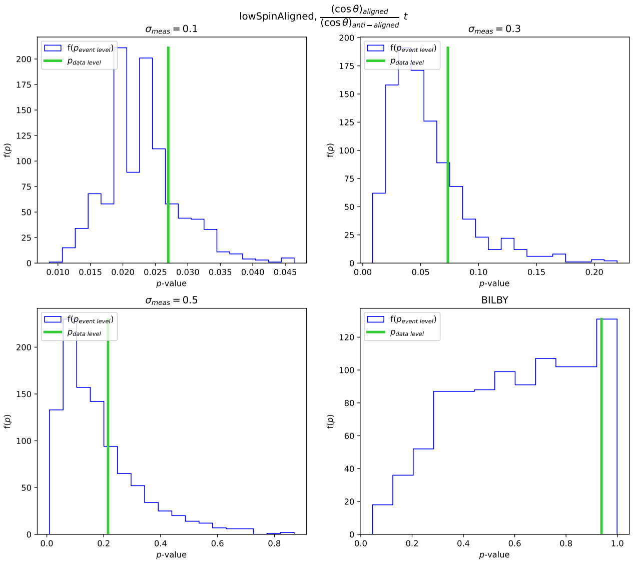

FIG. 7. The distribution of p-values for each of the traditional observed catalogs is shown by the blue histogram. The data-level PPC process produces observed catalogs that are identical, resulting in only one p-value, shown by the green line. The incompatibility of the model with the observed data is visually apparent for the Gaussian posteriors, and the data-level PPC is not more discerning than the traditional PPC. For the Bilby posteriors, the traditional PPC p-values become more uniformly distributed, which is a characteristic of a good model. The data-level p-value falls close to 1, indicating a model that severely underpredicts the observed test statistic.

p-values for the data-level PPCs are also shown. All of the observed catalogs from the data-level PPC are identical, so they each have the same p-value, shown by a green line. Even as measurement uncertainties increase and the traditional PPC p-values approach uniformity, the data-level p-value remains low, meaning data-level PPCs are better at showing model misspecification with uncertain data.

Fig. 6 shows the distribution of t for the traditional PPC when the test statistic is defined as the ***fraction of BBHs with aligned spins*** $(0.33 < \cos\theta \leq 1)$ ***divided by the fraction of those with anti-aligned spins*** $(-1 \geq \cos\theta < -0.33)$. For this test statistic, the data-level PPC p-value is not consistently lower than the traditional PPC p-value, as shown in Fig. 7. For $\sigma_{\rm meas} = 0.1,\ 0.3,\ 0.5$, the data-level p-values are on the tail of the traditional PPC p-value distributions and lie closer toward $p = 0.5$. For these measurement uncertainties, the data-level PPC is not better at identifying the bad model. For the Bilby posteriors, the data-level p-value is close to 1, meaning the PPC recognizes the bad model, while the traditional PPC p-value distribution is more uniform.

The results for the use of ***standard deviation*** as a test statistic is shown in Fig. 8. The distributions of t for the Gaussian posteriors with $\sigma_{\rm meas} = 0.1, 0.3$ showed little to no overlap, and both the traditional and data-level p-values clearly demonstrated the inaccuracy of model. For this reason, only the $\sigma_{\rm meas} = 0.5$ and Bilby posteriors are shown. For the $\sigma_{\rm meas} = 0.5$ case, the incompatibility of the model is apparent from the traditional PPC p-value distribution, which peaks around $p \approx 1$. The data-level PPC, which also falls around $p \approx 1$, is not more revealing. For the Bilby posterior, however, the traditional PPC p-values are widely distributed, implying consistency between the predicted and observed catalogs. Only the data-level p-value demonstrates the inaccuracy of the model and its tendency to overpredict $t_{\rm obs}$.

Fig. 9 shows the results with the test statistic defined as ***the fraction of BBHs with aligned spins***$(0.33 < \cos\theta \leq 1)$ ***divided by the fraction with small to zero spins*** $(-0.33 < \cos\theta < 0.33)$. Again, only the $\sigma_{\rm meas} = 0.5$ and Bilby posteriors are shown, and the traditional and data-level PPCs performed equally well for the $\sigma_{\rm meas} = 0.1, 0.3$ posteriors. Both the traditional and data-level p-values indicate the model inaccuracy for the $\sigma_{\rm meas} = 0.5$ posterior, For the Bilby posterior, neither show model misspecification. The traditional PPC p-values are widely distributed, and the data-level p-value falls around $p \approx 0.55$, suggesting a good fit between the model predictions and the observed data. Using this definition of t, neither the traditional nor the data-level PPCs are able to accurately assess the inaccurate model.

Across all definitions of t, the traditional distributions of $f(t_{\rm pred})$ and $f(t_{\rm obs})$ become closely aligned for the Bilby posterior due to the individual event reweighting step, leading to p-values that falsely indicate consistency between the observed data and the data predicted by the model. For most measurement uncertainties and choices of t, the data-level p-value is more reflective of the model misspecification because it is farther from $p = 0.5$. As we move forward, we aim to find predictive checks whose distribution of p-values skews more toward $p = 0$ or $p = 1$ for the test statistics that are highly distinguished between the population model and the true population.

IV. FUTURE OF RESEARCH: ALTERNATIVE POSTERIOR PREDICTIVE CHECKS

The objective of this project is to determine whether alternate 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. Partial Posterior Predictive Checks

Calculating the predictive posterior and evaluating the p-value with the same data can lead to a non-representative p-value. Partial PPCs address this short-coming of posterior predictive checking by avoiding the double-use of data. The partial PPC method does use the observed data to calculate the p-value, but only uses information not present in $t_{\rm obs}$ when training the prior. This eliminates one degree of freedom from the predictive posterior in order to glean the characteristics of the distribution that are not captured by the chosen $t_{\rm obs}$. The conditional distribution $f(x_{\rm obs}|t_{\rm obs}, \theta)$ is used as the likelihood to determine the posterior distribution $\pi_{\rm ppp}$ of event parameters θ,

$$\pi_{\rm ppp}(\theta|x_{\rm obs}\backslash t_{\rm obs}) \propto \mathcal{L}(x_{\rm obs}|t_{\rm obs}, \theta)\pi(\theta) \tag{4}$$

$$\propto \frac{\mathcal{L}(x_{\rm obs}|\theta)\pi(\theta)}{\mathcal{L}(t_{\rm obs}|\theta)}, \tag{5}$$

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

$$p_{\rm ppp}(t|x_{\rm obs}\backslash t_{\rm obs}) = \int \mathcal{L}(t|\theta)\pi(\theta|x_{\rm obs}\backslash t_{\rm obs})d\theta \tag{6}$$

The new p-value then takes the form

$$p = \Pr{}^{p_{\rm ppp}(t|x_{\rm obs}\backslash t_{\rm obs})}(t(\mathbf{x}) \geq t(\mathbf{x}_{\rm obs})). \tag{7}$$

The use of $f(x_{\rm obs}|t_{\rm obs}, \cos\theta)$ to define the likelihood instead of $f(x_{\rm obs}|\cos\theta)$ evades the double-use of data that occurs in PPCs when constructing the predictive posterior and then evaluating the p-value [18, 19].

To implement the partial PPC, we use the following steps:

1. Generate one observed trace using the traditional predictive posterior and calculate $t_{\rm obs}$.

2. Draw a set of hyperparameters Λ and reweight the injected population according to Λ and the proposed population model.

3. Draw a set of parameters from the reweighted injected population.

4. Repeat steps 2-3 until 70 predicted events have been drawn.

5. Calculate $t_{\rm pred}$ from the 70 predicted events. If $t_{\rm pred} = t_{\rm obs}$, this is the final predicted catalog. If $t_{\rm pred} \neq t_{\rm obs}$, repeat steps 2-4.

We repeat this process for each catalog.

B. Split Predictive Checks

The split predictive check (SPC) similarly aims to avoid repeated use of data. Rather than conditioning the prior for θ on the influence of $t_{\rm obs}$ (as in partial PPCs), however, split predictive checks partition the data into two disjoint subsets from the start. With a single split of data $x_{\rm obs} = x_a + x_b$, the method uses different subsets when training the posterior and when determining the p-value. The posterior distribution for x under the null model

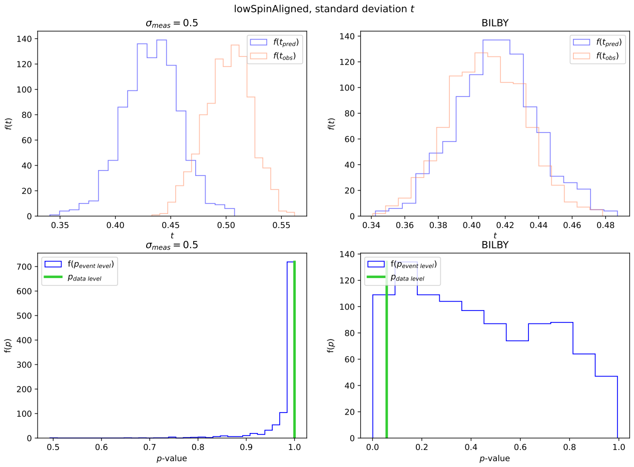

FIG. 8. **First row**: The observed and predicted distributions of the test statistic when defined as the standard deviation. The distribution of $t_{\rm obs}$ is shown by the pink histogram and the distribution of $t_{\rm pred}$ is shown by the blue histogram. Only the $\sigma_{\rm meas} = 0.5$ and the Bilby posteriors are shown; the t distributions from the $\sigma_{\rm meas} = 0.1, 0.3$ posteriors showed little to no overlap. **Second row**: The corresponding p-value distributions from the traditional observed catalogs (shown by the blue histogram) and the p-values for the data-level observed catalogs (shown by the green line). The observed catalogs from the data-level PPCs are identical, so only one p-value is plotted. For the $\sigma_{\rm meas} = 0.5$ posterior, the model misspecification is clear as a result of the t distributions only overlapping on the tails of each distribution. For the Bilby posterior, the significant overlap of the t distributions after reweighting individual events makes the model mismatch less apparent, causing the traditional PPC p-values to be distributed more uniformly. The data-level p-value, however, is low, indicating a poor model and a tendency to overpredict the observed standard deviation.

$$p_{\rm SPC}(x|x_a) = \int \mathcal{L}(x|\theta)\pi(\theta|x_a)d\theta \tag{8}$$

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

$$p = \Pr{}^{m_{\rm SPC}(x|x_a)}(t(x_b) \geq t(x)). \tag{9}$$

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 [20].

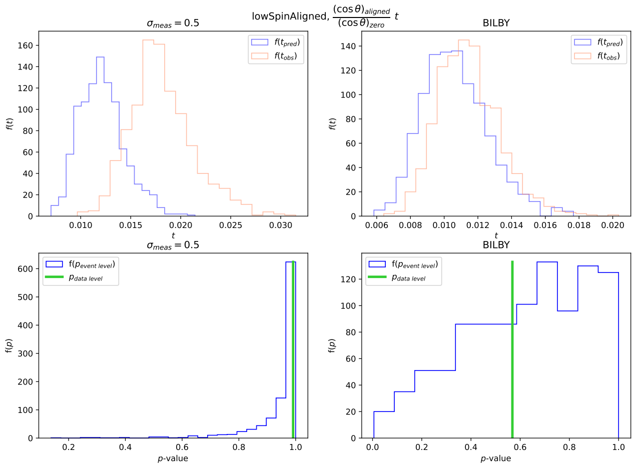

FIG. 9. **First row**: The observed and predicted distributions of the test statistic when defined as the number of aligned spins ($0.33 < \cos\theta \leq 1$) divided by the number of approximately zero spins ($-0.33 < \cos\theta < 0.33$) from the traditional PPC. The distribution of $t_{\rm obs}$ is shown by the pink histogram and the distribution of $t_{\rm pred}$ is shown by the blue histogram. Only the $\sigma_{\rm meas} = 0.5$ and the Bilby posteriors are shown; the t distributions from the $\sigma_{\rm meas} = 0.1, 0.3$ posteriors showed little to no overlap. **Second row**: The corresponding p-value distributions from the traditional observed catalogs (shown by the blue histogram) and the p-values for the data-level observed catalogs (shown by the green line). The observed catalogs from the data-level PPCs are identical, so only one p-value is plotted. For the $\sigma_{\rm meas} = 0.5$ posterior, the model misspecification is clear as a result of the t distributions overlapping only on the tails of each distribution. For the Bilby posterior, the significant overlap of the t distributions after reweighting individual events makes the model mismatch less apparent, causing the traditional PPC p-values to be distributed more uniformly. The data-level p-value is close to $p = 0.5$, suggesting a good model. Using this test statistic, the data-level PPC p-value is not able to reflect the model misspecification, nor is the traditional PPC.

In practice, this method is similar to the traditional PPC, but we perform a weighted hyperparameter draw rather than a random one. The weights for each hyperparameter are calculated by summing the likelihood of the hyperparameter given by our population model over a subset of individual events.

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 10 and 11.

Appendix B: 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.

[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. L. S. Collaboration, the Virgo Collaboration, and the KAGRA Collaboration, The population of merging compact binaries inferred using gravitational waves through gwtc-3 (2022), arXiv:2111.03634 [astro-ph.HE].

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

[7] S. Galaudage, C. Talbot, T. Nagar, D. Jain, E. Thrane, and I. Mandel, The Astrophysical Journal Letters **921**, L15 (2021).

[8] J. Fuller and L. Ma, The Astrophysical Journal Letters **881**, L1 (2019).

[9] D. Gerosa and M. Fishbach, Nature Astronomy **5**, 749–760 (2021).

[10] 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).

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

[12] 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].

[13] G. Ashton, M. Hübner, P. D. Lasky, C. Talbot, K. Ackley, S. Biscoveanu, Q. Chu, A. Divakarla, P. J. Easter, B. Goncharov, F. H. Vivanco, J. Harms, M. E. Lower, G. D. Meadors, D. Melchor, E. Payne, M. D. Pitkin,

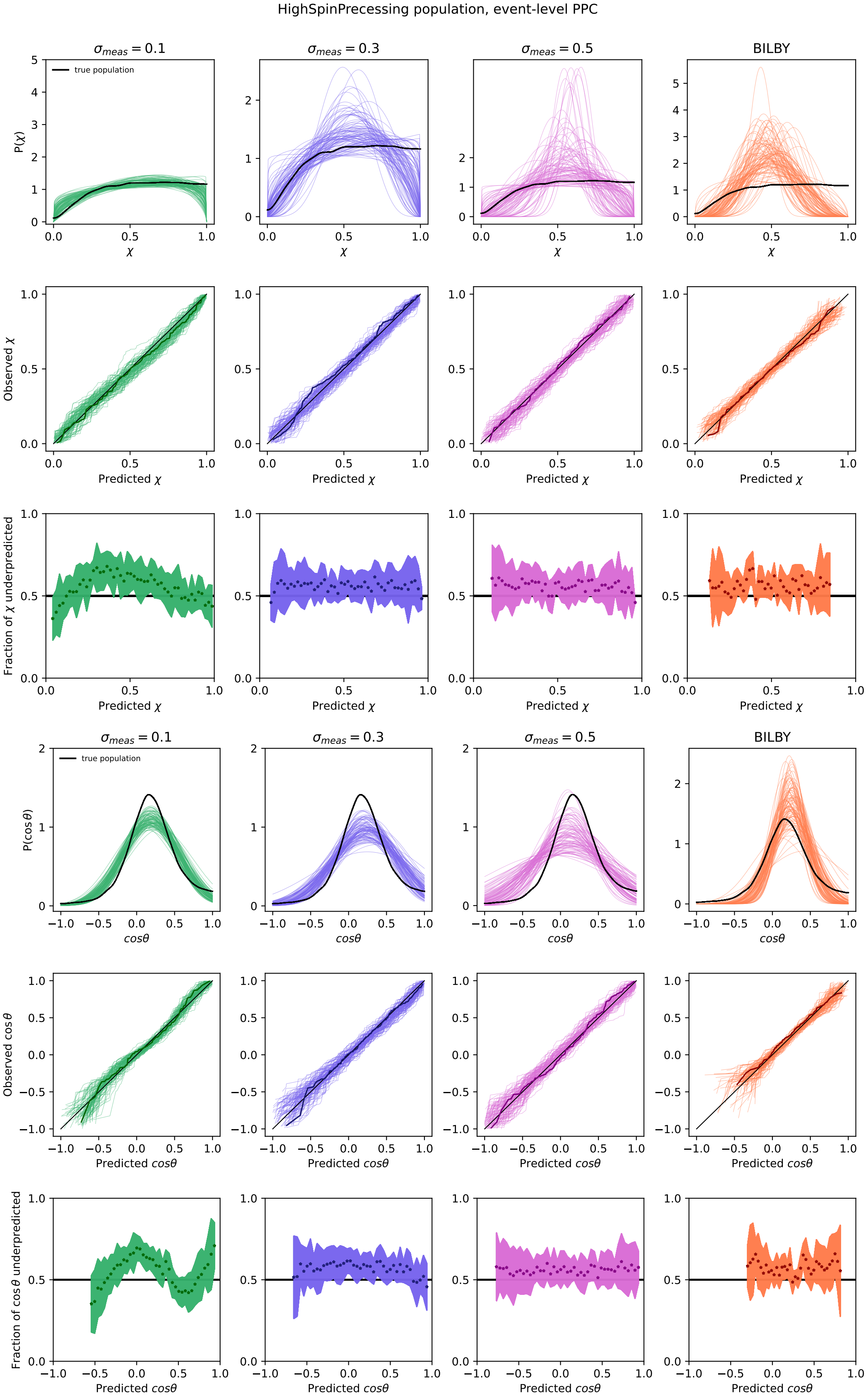

FIG. 10. Results of population inference using posterior predictive draws for χ and $\cos\theta$ of the HighSpinPrecessing simulated population.

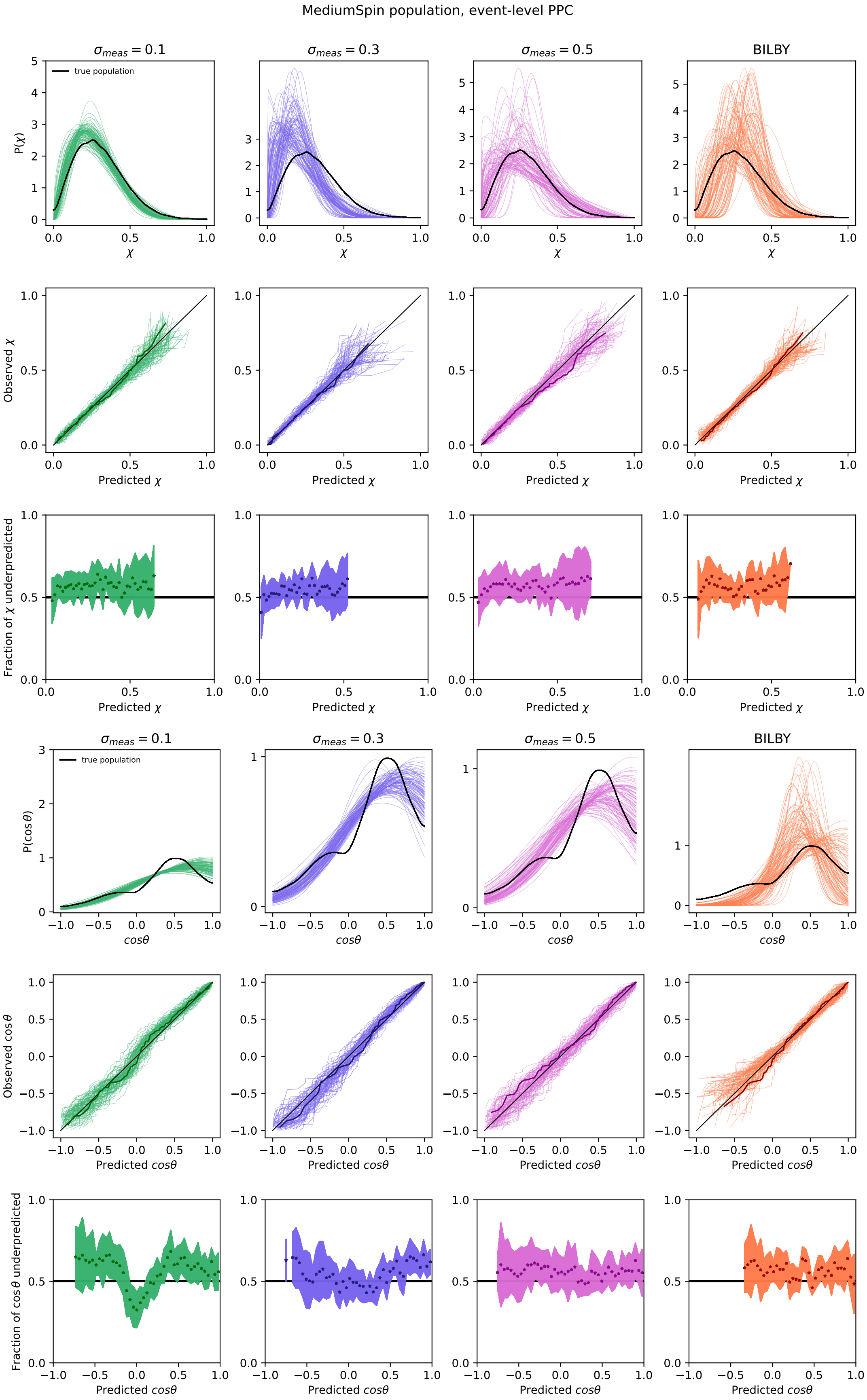

FIG. 11. Results of population inference using posterior predictive draws for χ and $\cos\theta$ of the MediumSpin simulated population.

J. Powell, N. Sarin, R. J. E. Smith, and E. Thrane, The Astrophysical Journal Supplement Series **241**, 27 (2019).

[14] G. Pratten, C. García-Quirós, M. Colleoni, A. Ramos-Buades, H. Estellés, M. Mateu-Lucena, R. Jaume, M. Haney, D. Keitel, J. E. Thompson, and S. Husa, Physical Review D **103**, 10.1103/physrevd.103.104056 (2021).

[15] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, Publications of the Astronomical Society of the Pacific **125**, 306–312 (2013).

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

[17] Colm Talbot, bilby.core.fisher.FisherMatrixPosteriorEstimator, https://lscsoft.docs.ligo.org/bilby/api/bilby.core.fisher.FisherMatrixPosteriorEstimator.html.

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

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

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