© 2023. The Author(s). Published by the American Astronomical Society.
OPEN ACCESS

Things That Might Go Bump in the Night: Assessing Structure in the Binary Black Hole Mass Spectrum

Amanda M. Farah[1], Bruce Edelman[2], Michael Zevin[3,4], Maya Fishbach[5], Jose María Ezquiaga[6], Ben Farr[2], and Daniel E. Holz[1,3,4]

[1] Department of Physics, University of Chicago, Chicago, IL 60637, USA; afarah@uchicago.edu
[2] Institute for Fundamental Science, Department of Physics, University of Oregon, Eugene, OR 97403, USA; bedelman@uoregon.edu
[3] Kavli Institute for Cosmological Physics, The University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA
[4] Enrico Fermi Institute, The University of Chicago, 933 East 56th Street, Chicago, IL 60637, USA
[5] Canadian Institute for Theoretical Astrophysics, David A. Dunlap Department of Astronomy and Astrophysics, and Department of Physics, 60 St George St, University of Toronto, Toronto, ON, M5S 3H8, Canada
[6] Niels Bohr International Academy, Niels Bohr Institute, Blegdamsvej 17, DK-2100 Copenhagen, Denmark

Received 2023 January 2; revised 2023 August 1; accepted 2023 August 1; published 2023 September 22

Abstract

Several features in the mass spectrum of merging binary black holes (BBHs) have been identified using data from the Third Gravitational Wave Transient Catalog (GWTC-3). These features are of particular interest as they may encode the uncertain mechanism of BBH formation. We assess if the features are statistically significant or the result of Poisson noise due to the finite number of observed events. We simulate catalogs of BBHs whose underlying distribution does not have the features of interest, apply the analysis previously performed on GWTC-3, and determine how often such features are spuriously found. We find that one of the features found in GWTC-3, the peak at $\sim 35\,M_\odot$, cannot be explained by Poisson noise alone: peaks as significant occur in 1.7% of catalogs generated from a featureless population. This peak is therefore likely to be of astrophysical origin. The data is suggestive of an additional significant peak at $\sim 10\,M_\odot$, though the exact location of this feature is not resolvable with current observations. Additional structure beyond a power law, such as the purported dip at $\sim 14\,M_\odot$, can be explained by Poisson noise. We also provide a publicly available package, `GWMockCat`, that creates simulated catalogs of BBH events with correlated measurement uncertainty and selection effects according to user-specified underlying distributions and detector sensitivities.

Unified Astronomy Thesaurus concepts: Astrophysical black holes (98); Stellar mass black holes (1611); Black holes (162); Gravitational waves (678); Gravitational wave sources (677); Gravitational wave astronomy (675); Bayesian statistics (1900); Hierarchical models (1925)

1. Introduction

Gravitational waves (GWs) from more than 70 mergers of compact objects have now been detected in the data of the LIGO (Aasi et al. 2015) and Virgo (Acernese et al. 2014) detectors. A cumulative catalog of these events and their properties has been produced by the LIGO–Virgo–KAGRA (LVK) collaborations. This collection of all detections to date is called the "Third Gravitational-Wave Transient Catalog" (GWTC-3; Abbott et al. 2021b), and has enabled several insights into the nature of gravity (Abbott et al. 2021c), the local expansion of the universe (Abbott et al. 2021d), and the population of GW sources (Abbott et al. 2021c).

The underlying population of GW sources holds information about the astrophysical processes that give rise to merging binaries of compact objects. The mass spectrum of binary black holes (BBHs), for example, encodes information about numerous physical processes underlying massive-star evolution, supernova physics, compact object formation, and binary interactions. For example, the presence or dearth of black holes with masses between ~ 2 and 5 $M_\odot$ (Özel et al. 2010; Farr et al. 2011; Fishbach et al. 2020a; Farah et al. 2022a) may unveil the maximum neutron star mass, the stability of mass transfer, and the timescales relevant for the engines that drive supernova explosions (e.g., Fryer et al. 2012; Mandel & Müller 2020; Zevin et al. 2020; Li et al. 2021; Patton et al. 2022; Siegel et al. 2023; van Son et al. 2022b). On the high-mass end, a sharp decrease in the mass spectrum for black holes with masses $\gtrsim 50\ M_\odot$ (Fishbach & Holz 2017; Edelman et al. 2021) would be a strong indication that the pair instability process is at play and limiting the core mass of massive stars (Fowler & Hoyle 1964; Barkat et al. 1967; Heger & Woosley 2002; Heger et al. 2003; Woosley & Heger 2015; Belczynski et al. 2016; Woosley 2017, 2019; Marchant et al. 2019; Renzo et al. 2020), with the location of the decrease in the differential merger rate acting to constrain relevant nuclear reaction rates (Farmer et al. 2020). Other overdensities and underdensities in the observed mass distribution (Tiwari & Fairhurst 2021; Edelman et al. 2022, 2022b; Tiwari 2022), as well as the evolution of the mass distribution with redshift (Fishbach et al. 2021; Karathanasis et al. 2023; van Son et al. 2022a, 2022b), will further inform the dominant BBH formation channels, binary evolution physics, and the metallicity evolution of the universe.

All of the parameters that are measurable from the signal of a binary merger can provide insight into formation mechanisms of merging binaries, especially when used in a population analysis (Stevenson et al. 2015; Zevin et al. 2017). However, the masses of the objects in the merging system are the best measured and span the largest dynamic range. Additionally, the mass distribution of compact objects can be used to measure

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

cosmological parameters using the "spectral siren" method, provided there is structure in the distribution beyond a boundless power law (Chernoff & Finn 1993; Messenger & Read 2012; Taylor et al. 2012; Farr et al. 2019; Abbott et al. 2021d; Ezquiaga & Holz 2021, 2022), such as edges, gaps, peaks, or changes in the power-law slope. Multiple features must be present to disentangle redshift evolution of the mass spectrum from cosmology, and more features further aid in breaking this degeneracy (Ezquiaga & Holz 2022). Therefore, considerable effort in the field of GW astronomy has gone toward understanding the mass distribution of GW sources. There are currently many more detected BBH mergers than binary neutron star (BNS) or neutron star–black hole (NSBH) mergers, so much of the activity has been on population properties of the BBH distribution, though the mass distribution of BNSs and NSBHs has also been been considered (Fishbach et al. 2020a; Landry & Read 2021; Biscoveanu et al. 2022b; Farah et al. 2022a; Ye & Fishbach 2022).

The mass distribution of merging BBHs is typically parameterized by the primary mass m_1, the larger of the two component masses in the binary, and the mass ratio $q = m_2/m_1$, the ratio of the less massive object's mass to the primary mass; though, other parameterizations are possible and valid (e.g., Fishbach & Holz 2020a; Tiwari & Fairhurst 2021; Farah et al. 2022a). The community has thus far gained a robust understanding of the large-scale features of the BBH mass distribution, and is just beginning to resolve its finer details. After the release of the First Gravitational-Wave Transient Catalog (Abbott et al. 2019b), minimum and maximum masses at $\sim 5\ M_\odot$ and $\sim 40\ M_\odot$ were identified in the BBH primary mass distribution, but it was not yet possible to distinguish between a uniform distribution and a power law between those two bounds (Fishbach & Holz 2017; Talbot & Thrane 2018; Abbott et al. 2019a). The Second Gravitational-Wave Transient Catalog (GWTC-2; Abbott et al. 2021d) brought dozens of additional events, and the BBH mass distribution was found to have a global maximum at $\sim 8\,M_\odot$ and an excess of BHs between $\sim 30\,M_\odot$–$40\,M_\odot$ followed by a steep, although not infinitely sharp, drop off in the rate at higher masses extending to $\sim 80\ M_\odot$ (instead of sharp cutoff at $\sim 40\ M_\odot$). At the time, there were not enough observations to determine whether the mass distribution had a local maximum at $\sim 35\ M_\odot$, represented by a Gaussian peak on top of a power law, or whether the steepening toward higher masses was better described as a break in the power law (Abbott et al. 2021a).

At the end of the third LIGO–Virgo observing run, the same two features at $\sim 8\,M_\odot$ and $\sim 35\ M_\odot$ remained, and the feature at $35\,M_\odot$ was classified as a peak rather than a break in the power law (Abbott et al. 2021c). Additionally, nonparametric (Mandel et al. 2017; Edelman et al. 2022b; Payne & Thrane 2023; Rinaldi & Del Pozzo 2022; Sadiq et al. 2022) and semiparametric (Edelman et al. 2022) analyses found robust evidence for an additional peak at $\sim 10\,M_\odot$, the same peak at $\sim 35\,M_\odot$, as well as modest evidence for a paucity of events near $\sim 14\,M_\odot$ (Abbott et al. 2021c). These features in the primary mass distribution correspond to similar ones in the chirp mass distribution, occurring at $\sim 9\,M_\odot$, $\sim 11\,M_\odot$, and $\sim 26\,M_\odot$, respectively (Tiwari & Fairhurst 2021; Tiwari 2022). The current picture of the BBH mass distribution is therefore a decreasing power law from low to high masses, with a global maximum at $m_1 \sim 10\,M_\odot$, a potential underdensity at $m_1 \sim 14\,M_\odot$, and an overdensity at $m_1 \sim 35\,M_\odot$. This can be seen in Figure 1, where we plot the results of fitting two parametric models and one semiparametric model to the BBHs in GWTC-3.

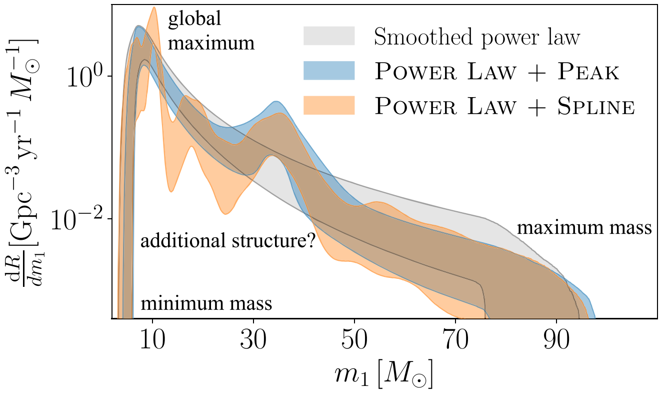

Figure 1. Distribution of primary BBH masses inferred using GWTC-3 and three different population models. The smoothed power-law model (gray) consists of a single power-law slope between a minimum and maximum mass, with the merger rate set to exactly zero outside of those bounds. It also includes a smoothing parameter at the low-mass end that allows for an offset between the minimum BH mass and the global maximum of the distribution. The Power Law + Peak model is similar to the smoothed power law, but also includes a Gaussian component. The Power Law + Spline model adds a cubic spline modulation to a smoothed power law to allow for additional substructure. We seek to determine if the perturbations beyond a power law found by Power Law + Spline and other semiparametric models can be explained by random associations in the data due to a finite number of observations, or if they are features of the true underlying distribution.

While the existence of this substructure in the current data set appears robust, its interpretation is less clear. Plausible explanations for this substructure include (i) Poisson noise, (ii) modeling systematics, or (iii) astrophysical signatures from one or several formation channels. We aim to disentangle the first two possibilities from the third using the Power Law + Spline model (Edelman et al. 2022), one of the semiparametric models used to identify the substructure reported in Abbott et al. (2021c).

Poisson noise would be caused by the fact that the fiducial BBH analysis in Abbott et al. (2021c) includes only 69 events over a mass range that spans more than an order of magnitude, so the observations may appear to be clumped at some masses even if the underlying distribution is smooth. We first determine if this explanation accounts for the data by simulating catalogs of BBHs whose underlying distribution does not have the features of interest, applying the analysis previously performed on GWTC-3, and determining how often such features are spuriously found. We develop several metrics comparing observations to simulated data in order to assess the statistical significance of the "bumps" in the primary mass distribution found by Abbott et al. (2021c), Edelman et al. (2022). All of the metrics derived in this work answer the same general question: how often do we infer the existence of a feature when analyzing observations of a true population *without* that feature? In this sense, these metrics are analogous to frequentist p-values, as lower values correspond to more significant features in the data. Readers familiar with gravitational-wave data analysis might find it useful to think of these metrics as false alarm rates (FARs) because they quantify how often noise resembles the observed signal.

A similar frequentist analysis on a large number of mock catalogs was performed by Sadiq et al. (2022) on the peak at $\sim 35\,M_\odot$ using an adaptive kernel density estimator (aKDE) to find features in samples drawn from featureless mass models, as well as from a model with a single peak. They account for selection effects, but not measurement uncertainty. They find that an aKDE is able to identify peaks in the data, and that the peak at $\sim 35\,M_\odot$ found in GWTC-2 is statistically significant within the aKDE model.

The second effect mentioned above, model systematics, could also plausibly cause spurious inference of features beyond a power law. It is potentially concerning that the models considered in Abbott et al. (2021c) that find peaks and troughs in the mass distribution are inherently "bumpy": both POWER LAW + PEAK (Talbot & Thrane 2018) and MULTI SOURCE employ a smoothed power law with a Gaussian component (Wysocki & O'Shaughnessy 2021), FLEXIBLE MIXTURES is a linear combination of Gaussian components, and POWER LAW + SPLINE employs a smoothed power law under a cubic spline modulation. The question is then whether these *bumpy* models can recover sharp features or if they instead create peaks and troughs that are morphologically dissimilar to the true distribution. This is most easily addressed by cross-checking with independent models such as BROKEN POWER LAW (Abbott et al. 2021a, 2021c) and the autoregressive model presented in Callister & Farr (2023).

Inaccuracies in the selection function are also known to cause systematic biases when inferring the underlying population (e.g., Malmquist 1922, 1925). These biases could, in principle, also cause an incorrect inference of structure in the astrophysical distribution of BBH masses. However, selection effects in GW detectors are remarkably well-characterized, so we expect this effect to be subdominant to Poisson uncertainty. As the number of events grows, so will our accuracy in the estimation of the selection function (Farr 2019; Essick & Farr 2022).

We provide posterior samples from our simulated catalogs in an accompanying data release (Farah et al. 2022b), and also provide a publicly available `python` package, `GWMockCat` (Farah et al. 2022c), to create similar samples according to user-defined populations.[7]

Section 2 provides a demonstrative example: it foregoes a full fit to the astrophysical population of sources, and compares the observed distribution of masses to possible observed distributions given an underlying power law in primary mass, (incorrectly) assuming no measurement uncertainty. This analysis suggests that the observed peak at $\sim 35\,M_\odot$ is statistically significant, but that all other features beyond a simple power law might be explainable by Poisson noise. This motivates a thorough study using a full hierarchical Bayesian analysis on simulated event posteriors, which we carry out in Section 3. Section 4 summarizes our conclusions and discusses their implications for the astrophysical origin of the GWs observed thus far by the LVK. Readers primarily interested in the significance of features in the mass distribution may wish to skip to Section 3.3, whereas those interested in using the package `GWMockCat` can find details in Appendices A and B.

2. Motivation

To construct a simple test of feature significance and motivate further study, we first avoid a fit to the mass distribution and instead consider the *observed* distribution of primary masses and its resemblance to one that would result from a simple power law. The observed population differs significantly from the astrophysical one, as current gravitational-wave detectors are subject to selection biases that favor the detection of closer and more massive systems, as well as measurement error that affects each system differently. We construct plausible observed mass distributions that could occur from detecting 69 BBHs whose astrophysical distribution is a featureless power law in primary mass. To do this, we use the samples provided by LIGO Scientific Collaboration et al. (2021a), which were created for sensitivity estimation for the LVK's GWTC-3 analysis. Each of these samples comes with a probability of being drawn from an assumed underlying distribution and a FAR assigned by each search used by the LVK. We can then reweight these samples to our desired population model (in this case, a power law in m_1, q, and z) using the draw probability, and apply the same FAR threshold used in Abbott et al. (2021c) to select "found injections." Of the $\sim 6 \times 10^4$ found injections, we resample to $N = 10^4$ independent sets of 69 draws each to directly compare to the observations.

We then histogram each set of these found injections, thereby obtaining a distribution of bin heights for our mock populations. Using several thousand realizations of found injection sets enables us to construct a null distribution of bin heights and characterize the effect of Poisson noise on the shape of the observed distribution. We compare these null distributions with the observed distribution of BBH masses in GWTC-3[8] by assuming the primary masses are measured perfectly and using the median a posteriori values of their primary masses as point estimates. The result is shown in the top panel of Figure 2, which plots the 90% credible interval on the observed null distributions, along with the distribution of median primary masses of GWTC-3's BBHs. For the null distributions, we consider two power-law spectral indices as representative examples: $\alpha = 2.7$, and $\alpha = 3.25$. These are chosen to represent a range of plausible values for the BBHs in GWTC-3: a power-law fit to GWTC-3 yields $\alpha = 2.98^{+0.16}_{-0.28}$, where the bounds represent 1σ deviations.

To obtain a more quantitative measure, we compare bin heights from the found injections, $h_{\rm inj}$, to the bin heights of observed events in GWTC-3, $h_{\rm GWTC-3}$, obtaining for each bin i the fraction of simulated bin heights that are lower than those of GWTC-3 BBHs. Explicitly,

$$r_h^i = \frac{1}{N}\sum_j^N \begin{cases} 1 & \text{if } h_{i,\rm inj}^j < h_{i,\rm GWTC-3} \\ 0 & \text{if } h_{i,\rm inj}^j \geqslant h_{i,\rm GWTC-3} \end{cases}, \qquad (1)$$

where the sum is over the $N = 10^4$ sets of found injections, and r_h is defined for each bin. For example, if $r_h = 0.95$ for a given bin, the observed distribution in that bin is larger than would be expected from a featureless power law 95% of the time. A value of r_h approaching unity corresponds to a *bump* in the

[7] The data release can be found at doi:10.5281/zenodo.7411991, and `GWMockCat` can be installed at https://git.ligo.org/amanda.farah/mock-PE.

[8] For all comparisons to real observations, we use the publicly available posterior samples for the GWTC-2.1 and GWTC-3 data releases (LIGO Scientific Collaboration & Virgo Collaboration 2022; LIGO Scientific Collaboration et al. 2021b, respectively). We use samples generated with the IMRPhenomXPHM waveform and a prior proportional to the square of the luminosity distances (i.e., the samples were not "cosmologically reweighted"). To make the most direct comparison with Abbott et al. (2021c), we keep events with secondary mass larger than $3\,M_\odot$ and FAR less than 1 yr^{-1}, resulting in 69 events.

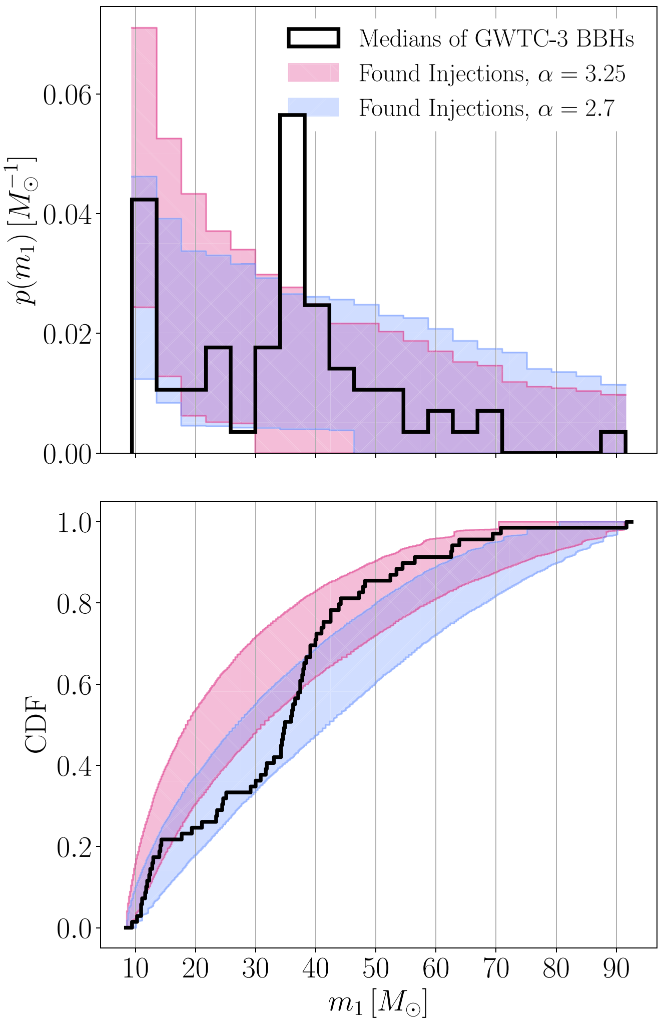

Figure 2. Observed source-frame primary mass distributions. Black solid lines contain the median a posteriori values for the binary black holes in GWTC-3. Pink and blue bands indicate the 90% credible interval on the *observed* distributions predicted from astrophysical distributions that are power laws in primary mass with spectral index $\alpha = 3.25$, and $\alpha = 2.7$, respectively. The top panel shows a histogram of observed primary masses. For GWTC-3's distribution to be consistent with the null distributions, we expect its bin heights in the top panel to be within the 90% credible intervals in 18 out of the 20 bins. The uncertainties in these predicted distributions are due only to Poisson noise resulting from a finite number of observations, rather than modeling uncertainty or uncertainty in parameter estimation. Therefore, the cumulative distribution functions in the bottom panel are similar to a conventional posterior predictive check, but with only one source of uncertainty. The large deviations of the black curve from the shaded bands in some regions indicate the difficulty that a single power law with Poisson shot noise has in fully explaining the observations. However, many of the apparent excursions from a power law are well-contained within the predicted bands.

observed mass distribution, and a value of r_h approaching zero is indicative of a *dip*.

Note that the comparison between the null distributions and GWTC-3 is occurring at each bin, rather than across all bins. We do this because the magnitude of Poisson noise depends on the value of m_1: since the underlying distribution is not uniform, fewer events are expected at very high m_1, and therefore, the relative standard deviation is larger. This is also a consequence of Eddington bias (Eddington 1913). Making comparisons at specific points in m_1 does not, however, properly correct for the look-elsewhere effect. We will address this effect in Section 3.

The three most significant values of r_h^i in the case of $\alpha = 3.25$ are $r_h^{15.6\,M_\odot} = 0.033$, $r_h^{27.9\,M_\odot} = 0.036$, $r_h^{36.1\,M_\odot} > 0.999$, where the superscripts indicate the centers of the bins at which r was calculated. This means that less than 0.1% of mock populations had more events near 36.1 $M_\odot$ than GWTC-3 does, 3.3% of mock populations had fewer events near 15.6 $M_\odot$ than GWTC-3, and at 27.9 $M_\odot$, 3.6% of mock populations had fewer events.

Repeating the exercise for $\alpha = 2.7$, we find the three most significant values of $r_h^i$to be $r_h^{40.2\,M_\odot} = 0.935$, $r_h^{27.9\,M_\odot} = 0.020$, $r_h^{36.1\,M_\odot} > 0.999$. The locations of the significant features differ when the assumed underlying distribution changes.[9] In either case, the bump at $\sim$35 $M_\odot$ is unlikely to be due to Poisson noise, but other features may be.

To avoid the need to arbitrarily choose bins, we additionally construct a cumulative distribution function (CDF) of the primary masses and compare it to the CDFs of the null distributions, shown in the bottom panel of Figure 2. This comparison is akin to a posterior predictive check in that it can highlight where the model fails to predict the data. Importantly, though, it differs from the conventional posterior predictive check because we have purposefully left out the effects of modeling uncertainty and measurement uncertainty in order to isolate the effects of Poisson noise. The prior distributions are therefore also not included, since each event is assumed to be measured with perfect accuracy.

If $\alpha = 3.25$, the null distributions are consistent with the data below $\sim$18 $M_\odot$ and above $\sim$35 $M_\odot$, but not between them, meaning that the $\sim$10 $M_\odot$ and $\sim$35 $M_\odot$ peaks can be explained by Poisson noise, but the underdensity between them could not be. On the other hand, if $\alpha = 2.7$, the null distributions are consistent with the data everywhere except for above $\sim$40 $M_\odot$, suggesting that, under this scenario, Poisson noise can explain all features except for the $\sim$35 $M_\odot$ peak.

For both spectral indices considered, two of the three features found by Abbott et al. (2021c) can be explained by Poisson noise from a finite number of observations. However, this does not mean that exactly two of the features are the result of Poisson noise, just that no more than two can be caused by the phenomenon. Additionally, it is not clear *which* of the features are more likely to have physical origin, as this method offers no quantitative way to determine which power-law slope is preferred.

Importantly, this methodology does not account for the effects of measurement error, which can cause significant biases near the edges of sharp distributions when not properly accounted for (Fishbach et al. 2020b). We therefore turn to a full hierarchical Bayesian analysis of simulated catalogs, which will allow us to fit for the population model parameters, take the measurement uncertainty into account, and directly compare to the metrics used in Abbott et al. (2021c).

3. Hierarchical Analysis and Results

We determine how often the features inferred in the mass distribution of BBHs would be spuriously found in data whose

[9] It is also possible to determine the existence of local minima or maxima in this observed distribution independently of the underlying power law. This can be done using a dip test for unbinned data (Hartigan & Hartigan 1985) or the minimum number of components required for a Gaussian mixture model (McLachlan & Peel 2000). However, features in the observed distribution would be difficult to disentangle from selection effects, so we recommend only applying these to the astrophysical distribution, as in Tiwari & Fairhurst (2021). Since our principal aim is to quantify the significance of excursions from a power law, we leave such tests for future work.

underlying distribution does not have those features. To do this, we construct a null distribution by simulating BBH observations that would occur if the underlying astrophysical distribution was a single power law with no substructure in a finite range. The procedure for creating synthetic BBH observations is described in Appendix A. Mock observations are combined with corresponding sensitivity estimates in a hierarchical Bayesian analysis, described in Loredo (2009), Mandel et al. (2019), and Thrane & Talbot (2019). We analyze these simulations in the same way as the BBHs in GWTC-3 to determine how often the features observed in GWTC-3 would be found from an underlying distribution without those features.

3.1. POWER LAW + SPLINE Mass Model

We use the POWER LAW + SPLINE semiparametric primary mass model as a flexible model that is easily capable of finding peaks and valleys in the mass distribution (Abbott et al. 2021c; Edelman et al. 2022). This model parameterizes perturbations or deviations from a simpler underlying distribution with flexible cubic spline functions. Specifically, given an underlying hyper-prior for primary mass, $p(m_1|\Lambda)$, the POWER LAW + SPLINE model describes the primary mass distribution as follows:

$$p_{\text{spline}}(m_1|\Lambda, \{m_i\}, \{f_i\}) \propto p(m_1|\Lambda)\exp(f(m_1|\{m_i\}, \{f_i\})), \tag{2}$$

where $f(m_1|\{m_i\}, \{f_i\})$ is the function describing the perturbations, which we model with a cubic spline function interpolated by introduced hyperparameters, $\{m_i\}$, the locations of spline knots in mass space, and $\{f_i\}$, the height of the perturbation function at each knot. This describes a semiparametric model as it includes a simple *parametric* component (the underlying distribution) in addition to a nonparametric component that models the perturbation around the simple description. For this study, we use the simplest primary mass model for the underlying description, which is the TRUNCATED model, describing a power law with sharp cutoffs at the lower and upper mass bounds (Fishbach & Holz 2017; Edelman et al. 2022). While this model has been shown to insufficiently describe the primary mass distribution, it captures the majority of the broadest features (Abbott et al. 2021a, 2021c).

To assess the significance of peaks or valleys found with the POWER LAW + SPLINE model, one can look at the posterior distribution of the perturbation heights as a function of mass. This tells us how far *off* the simple power-law description is from accounting for the data. Specifically, we can find what percentile $f(m_1) = 0$ falls in the posterior distribution as a function of mass. For data exactly distributed as a power law (the underlying population), the inferred perturbation function should be symmetric about 0 with widths determined by the prior distributions on the knot heights and the number of observed events. At masses where the percentile of zero perturbation approaches 100% (0%), we can say there is an overdensity (or underdensity) of events at these masses, compared to the underlying power-law distribution. This is identical to the analysis done by Abbott et al. (2021c), who use the percentile at which the perturbation function excludes zero at a given location as a metric for how significant a feature is at that location.

3.2. Metrics of Feature Significance

As described in Section 3.1, the POWER LAW + SPLINE model makes use of a perturbation function constructed from cubic splines. The height of the perturbation function, $f(m_1)$, at a point in primary mass, m_1, is then a direct measure of the deviation from a power law at that point. We can determine how often one would find spurious evidence for substructure by simulating catalogs from a power law, fitting them with the POWER LAW + SPLINE model, and examining the resulting perturbation function.

If the mock catalogs produce perturbation functions with similar amplitudes to those seen for GWTC-3, the structure in the GWTC-3 fit might be described by Poisson noise. On the other hand, if the perturbation functions produced by fits to the mock catalogs are always lower in amplitude to that of the GWTC-3 fit, the structure in the GWTC-3 data is likely to be present in the underlying distribution.

For a given mock catalog, we find the m_1 value where the median a posteriori value of the perturbation function is maximal. We obtain the posterior distribution of perturbation function amplitudes at that location, $g(f_{\max})$. We repeat this for all mock catalogs, obtaining a set of maximal perturbation function distributions, $\{g_j(f_{\max})\}$. These are plotted in light gray on the left panels of Figure 4. The locations of the three maximal perturbation function amplitudes in the GWTC-3 fit are, from least to most Bayesian significance, $13.8\,M_\odot$, $10.3\,M_\odot$, and $35.7\,M_\odot$. The posterior distributions of perturbation function heights at these locations are $g_{\text{GWTC}-3}(f(13.8\,M_\odot))$, $g_{\text{GWTC}-3}(f(10.3\,M_\odot))$, and $g_{\text{GWTC}-3}(f(35.7\,M_\odot))$, and are plotted in orange in the left panels of Figure 4. The amplitude of the perturbation function at $13.5\,M_\odot$ is negative (i.e., it is a dip rather than a bump), so we flip its distribution about zero for more direct comparison. The same is done for all $g(f_{\max})$ whose medians are negative, as the perturbation function's prior is symmetric about zero.

3.3. Simulation Study

To determine whether the features in the mass spectrum of GWTC-3 BBHs are the result of Poisson noise of a finite number of observations drawn from a featureless power law, we compare POWER LAW + SPLINE fits using the GWTC-3 catalog and 300 mock catalogs generated from a *featureless* power law. The mock catalogs considered in this section are all generated from the same underlying distribution: a truncated power law in primary mass, mass ratio, and redshift, with a smoothing at low component masses to ensure the peak of the mass distribution is not in the same location as the minimum mass. The explicit form of the mock catalogs' population model, including values of all of its hyperparameters, can be found in Appendix B.

Full parameter estimation was not performed on each event in each mock catalog; instead, we use prescriptions for generating event posteriors that reproduce the correlations between an event's parameters, as well as the typical uncertainties seen in GWTC-3. We show in Appendix C that the prescriptions used are sufficient to reproduce population analyses such as the ones scrutinized in this work. In order for our comparisons to be consistent between featureless mock catalogs and GWTC-3, we recreate GWTC-3's event posteriors with the same prescriptions as were used for the mock catalogs, perform a population analysis on those, and use the resulting perturbation function for all comparisons to mock catalogs. Our

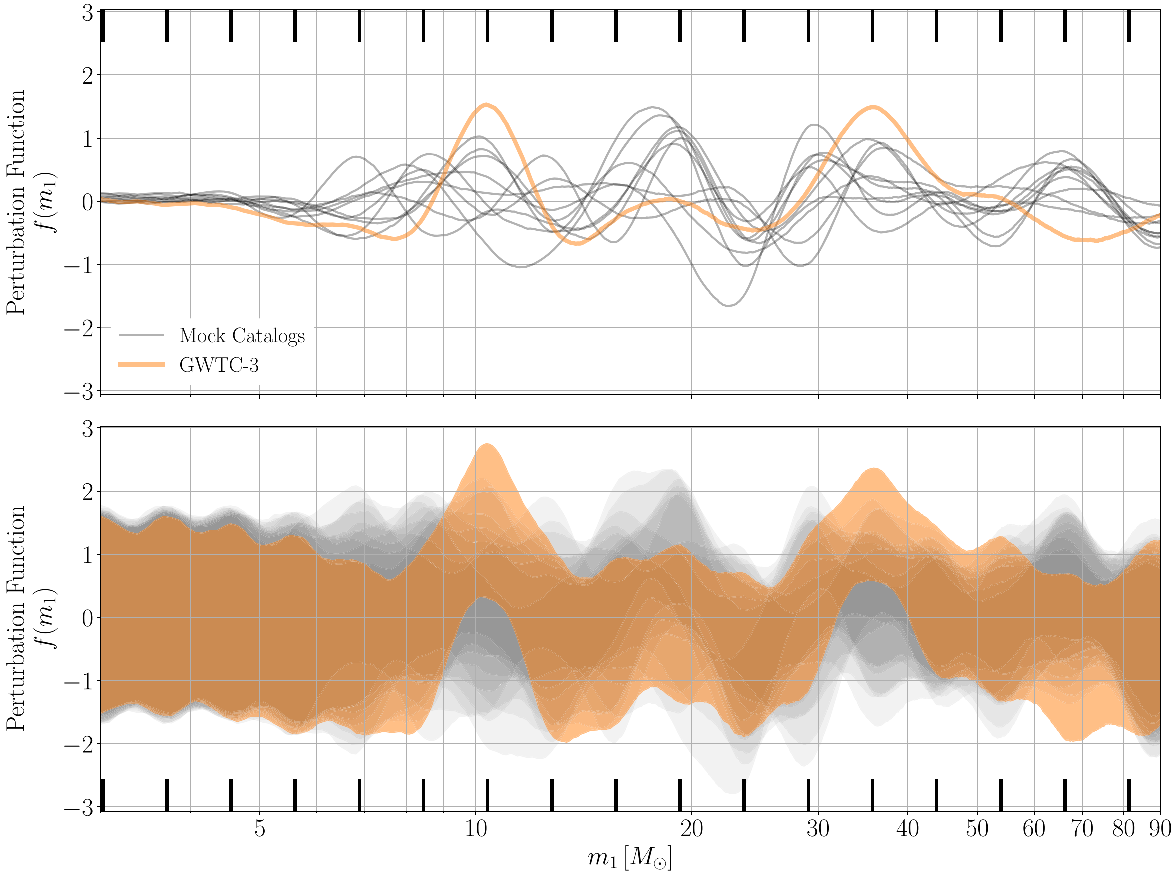

Figure 3. Median (top panel) and 90% credible interval (bottom panel) of the perturbation function resulting from the POWER LAW + SPLINE fit to the primary masses in GWTC-3 (orange) and in 10 mock catalogs (gray). The perturbation function multiplies a smoothed power law in primary mass to add modulations to an otherwise monotonic distribution, making it a direct measure of deviations from a power law. It is a cubic spline with knots fixed at the locations indicated by the black vertical tick marks. The prior on the perturbation heights is the unit normal distribution, as can be seen below ∼5 $M_\odot$ where there are no detections to constrain the likelihood and the posterior reverts to the prior. The perturbation function corresponding to GWTC-3 events appears large in amplitude in three locations: ∼10 $M_\odot$, ∼14 $M_\odot$, and ∼35 $M_\odot$. While the medians of the perturbation function at these distributions are comparable in amplitude, the posterior distribution at ∼35 $M_\odot$ (∼14 $M_\odot$) is the most (least) tightly constrained.

population reanalysis of the GWTC-3 events with mock posteriors appears consistent with the full analysis presented by the LVK in Abbott et al. (2021c). The lines labeled "GWTC-3" or depicted in orange in Figures 3, 4, and 5 refer to the analysis done on the recreated version of GWTC-3. All analyses presented here were repeated using the original LVK-released parameter estimation samples in Appendix C, and the same qualitative conclusions were reached, although with slightly more statistical significance.

Despite knowing the parameters of the underlying population for the mock catalogs, we allow all hyperparameters to vary when fitting POWER LAW + SPLINE to the mock catalogs. The resulting perturbation functions are shown in Figure 3 for 10 randomly chosen mock catalogs and GWTC-3. The perturbation functions deviate from their prior distribution in the mass range where detections exist (above ∼5 $M_\odot$ and below ∼85 $M_\odot$), even in the case of mock catalogs. This means that the perturbation functions are informed by the mock data despite the mock data not inherently requiring a deviation from a power law. The question still remains whether the perturbation function heights inferred from mock catalogs with no substructure are larger than those inferred from GWTC-3. While nonzero values of the perturbation function are common in the 10 mock catalog fits shown in Figure 3, only a few amplitudes appear comparable in height to the three largest amplitudes of the GWTC-3 perturbation function.

To verify this, we isolate the largest amplitude perturbations for all 300 mock catalog fits and compare them to the three largest amplitude perturbations for the GWTC-3 fit. These are plotted in the leftmost panels of Figure 4. The light gray curves are the posterior distributions of largest perturbation function amplitudes $\{g_j(f_{\max})\}$ for each simulated catalog j. These appear to have the same general shape as one another, although with noticeable scatter. The orange curves in each panel are the posterior distributions of GWTC-3's perturbation function $g_{\text{GWTC}-3}(f(m_1))$ at its three maximal locations: $m_1 = 13.8\,M_\odot$, $10.3\,M_\odot$, and $35.7\,M_\odot$.

The distribution for the ∼14 $M_\odot$ dip appears qualitatively similar to that of the simulated catalogs, the ∼10 $M_\odot$ peak appears to be slightly shifted with respect to most of the simulated catalogs but still within their range, and the ∼35 $M_\odot$ peak is noticeably shifted toward higher values relative to the bulk of the simulated catalog distributions. This suggests that the ∼35 $M_\odot$ peak is unlikely to be the result of Poisson noise or modeling systematics, while other features could plausibly be explained by those effects.

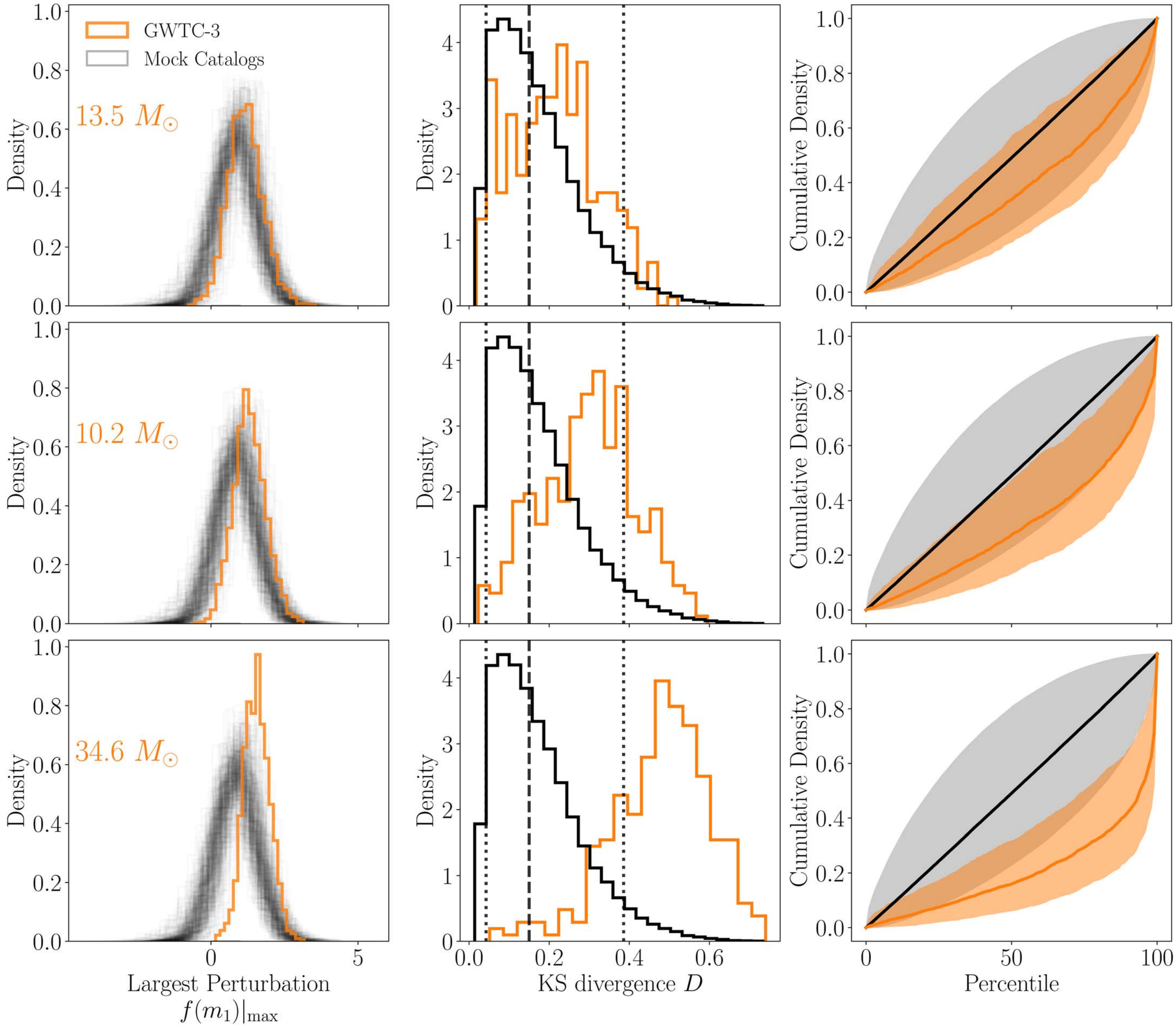

Figure 4. Three largest deviations from a power law observed in GWTC-3 compared to mock catalogs. Left column: the posterior distribution of perturbation function heights at the location where the posterior distribution is maximal for mock catalogs (light gray) and GWTC-3 (solid orange). Middle column: null distribution (black) and GWTC-3 distribution (orange) of Kolmogorov–Smirnov (KS) divergences between the individual distributions in the left column. Smaller values of the KS divergence indicate more similar distributions. Right column: null distribution (black) and GWTC-3 distribution (orange) of percentiles. Large deviations from the diagonal indicate a more significant rightward shift of the GWTC-3 distribution relative to the mock catalogs. Each row corresponds to a different local extremum for GWTC-3: $m_1 = 13.8\,M_\odot$ (top), $m_1 = 10.3\,M_\odot$ (middle), and $m_1 = 35.7\,M_\odot$ (bottom), while the global extrema for each mock catalog are shown in all rows, along with the aggregated distribution across all mock catalogs (solid black). The $\sim 35\,M_\odot$ peak is an outlier with respect to both the KS and percentile statistics, but the other two features are more ambiguous.

3.3.1. Maximum Perturbation Amplitude

To obtain a more quantitative measure, we derive several metrics from the distributions of maximal perturbation function amplitudes. The first uses the Kolmogorov–Smirnov (KS) test: we compute the KS divergence D between each of the $\{g_j(f_{\max})\}$ distributions to obtain a null distribution of KS divergences, shown in the solid black curve in the middle column of Figure 4. We then perform a KS test between the $\{g_j(f_{\max})\}$ distributions and $g_{\mathrm{GWTC-3}}(f(m_1))$ and obtain the orange curves in the middle column of Figure 4. From this, we find that the KS divergences for mock catalogs are larger than those of GWTC-3 20%, 11%, and 7% of the time for the $14\,M_\odot$, $10\,M_\odot$, and $35\,M_\odot$ features, respectively. This means, for example, that mock catalogs can produce perturbation function posteriors as tall as the one inferred from GWTC-3 at $\sim 35\,M_\odot$ only 7% of the time. Written in terms of $g(f)$, $g_{\mathrm{GWTC-3}}(f(14\,M_\odot)) \neq g_j(f_{\max})$ to 20%, $g_{\mathrm{GWTC-3}}(f(10\,M_\odot)) \neq g_j(f_{\max})$ to 11%, and $g_{\mathrm{GWTC-3}}(f(35\,M_\odot)) \neq g_j(f_{\max})$ to 7%. Although none of these percentages are convincingly small, this indicates that the orange histograms are more statistically distinct from the black histograms in the case of the $\sim 35\,M_\odot$ peak than they are in the cases of the features at $10\,M_\odot$ and $14\,M_\odot$.

The second metric is obtained by quantifying the shift of $g_{\mathrm{GWTC-3}}(f(m_1))$ relative to the set of $\{g_j(f_{\max})\}$. For each point

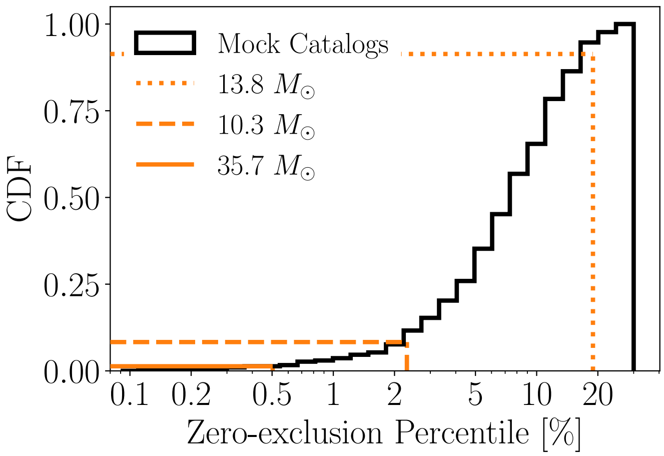

Figure 5. Percentile at which the posterior distribution of the perturbation function excludes zero for GWTC-3 (orange vertical lines) and catalogs drawn from a featureless distribution (black histogram). For GWTC-3, we evaluate the perturbation function's posterior distribution at primary mass (m_1) values of 13.8 $M_\odot$ (dotted), 10.3 $M_\odot$ (dashed), and 35.7 $M_\odot$ (solid). For mock catalogs, we find the primary mass value at which the perturbation function is maximal and evaluate its posterior distribution there. The values reported here are the percentage of the posterior distribution that is greater than zero at those values in m_1. The 13.5 $M_\odot$ feature excludes zero to a level comparable to some of the mock catalogs, but the other two features exclude zero to a level not reproducible by any mock catalogs.

in $g_{\mathrm{GWTC-3}}(f(m_1))$, we calculate the percentile in which it lies in each of the $\{g_j(f_{\max})\}$, obtaining the orange bands in the rightmost panels of Figure 4. For comparison, we do the same for each of the $\{g_j(f_{\max})\}$ relative to each other, constructing the gray bands in the rightmost panels of Figure 4. We then take the mean of the set of light orange bands and light black bands to obtain the solid orange and solid black curves, respectively. The black bands serve as null distributions, so large deviations from those indicate significant shifts. We observe a large deviation for the $\sim$35 $M_\odot$ peak, a moderate deviation for the $\sim$10 $M_\odot$ peak, and only a slight deviation for the $\sim$14 $M_\odot$ dip. Quantitatively, $g_{\mathrm{GWTC-3}}(f(35\,M_\odot)) \geqslant g_j(f_{\max})$ to $83^{+17}_{-69}\%$ (90% credible interval), meaning that the $\sim$35 $M_\odot$ peak lies in the 83^{+17}_{-69}rd percentile of the mock catalogs' largest perturbation heights. For the other features, $g_{\mathrm{GWTC-3}}(f(10\,M_\odot)) \geqslant g_j(f_{\max})$ to $74^{+25}_{-60}\%$, and $g_{\mathrm{GWTC-3}}(f(14\,M_\odot)) \geqslant g_j(f_{\max})$ to $34^{+52}_{-32}\%$. In comparison, the corresponding statistic for the null distributions is $g_j(f_{\max}) \geqslant g_i(f_{\max})$ to $50^{+47}_{-46}\%$.

It is not possible to draw firm conclusions from these large uncertainties. However, the central values indicate that the $\sim$35 $M_\odot$ peak is noticeably shifted relative to the mock catalogs' perturbation functions, the $\sim$10 $M_\odot$ peak is moderately shifted, and the $\sim$14 $M_\odot$ dip even has a slightly lower amplitude than the maximum perturbation functions typical of mock catalogs. We repeat this analysis with an Anderson–Darling test rather than a KS test and find similar results.

3.3.2. Inconsistency with a Power Law

The final metric we consider is inspired by the statistic presented in Abbott et al. (2021c), which states that "the inferred perturbation $f(m_1)$ strongly disfavors zero at both the 10 $M_\odot$ and 35 $M_\odot$ peak." We therefore turn from considering the full distribution of perturbation function heights at a given location to the percentile at which it excludes zero. A perturbation function amplitude of zero is a useful reference point for several reasons. The most intuitive is that it causes the population model to behave like a featureless power law, so a posterior that excludes zero to high credibility indicates an inconsistency with a power law. Zero is also the mean of the prior predictive distribution for the perturbation function: the prior allows for equal upwards and downwards fluctuations, symmetric about zero perturbation. Similarly, a vanishing perturbation function amplitude is the state to which we expect the posterior predictive distribution to asymptote in the limit of infinite detections from an underlying power-law distribution. We therefore plot the percentile at which each mock catalog excludes zero perturbation in Figure 5.

We then calculate how often a simulated catalog's perturbation function excludes zero to the same credibility as that of GWTC-3. This is the same as finding the point along the y-axis of Figure 5 at which each of the vertical orange lines hits the CDF. 1.7%, 10.0%, and 92.7% of the $\{g_j(f_{\max})\}$ exclude zero to the same percentile as $g_{\mathrm{GWTC-3}}(f(35\,M_\odot))$, $g_{\mathrm{GWTC-3}}(f(10\,M_\odot))$, and $g_{\mathrm{GWTC-3}}(f(14\,M_\odot))$, respectively. The fact that, for example, $g_{\mathrm{GWTC-3}}(f(14\,M_\odot)) < 0$ to 20.7% but 92.7% of mock catalogs have a similar or smaller statistical excursion is due in part to the difference between Bayesian credible intervals and frequentist p-values, and because our metric corrects for the look-elsewhere effect by comparing GWTC-3's perturbation function at specific locations to all possible locations in the mock catalogs.

Combined with the metrics presented in Section 3.3.1, the results above lead us to conclude that the peak at $\sim$35 $M_\odot$ is difficult to reproduce with featureless catalogs, but it is possible that the dip at $\sim$14 $M_\odot$ is just a large fluctuation rather than a astrophysical feature. The peak at $\sim$10 $M_\odot$ is difficult to reproduce with featureless catalogs; though, it is easier to reproduce than the $\sim$35 $M_\odot$ peak. We discuss the interpretation of this feature in more detail in Section 4.

In summary, featureless catalogs can sometimes produce features as tall as the $\sim$10 $M_\odot$ peak, and they can sometimes produce perturbations constrained away from zero with the same credibility. The dip at $\sim$14 $M_\odot$ could be a Poisson fluctuation because fits to featureless catalogs can easily produce perturbations as large, and as credibly constrained away from zero perturbation. The peak at $\sim$35 $M_\odot$ is difficult to reproduce by mock catalogs in any way: its perturbation amplitude is too large and too credibly constrained away from zero.

The fact that we find one of the features explainable by Poisson noise is consistent with Section 2, which suggests that up to two of the excursions from a power law can be explained by Poisson fluctuations. Our conclusions are also in broad agreement with those presented in Abbott et al. (2021c), as they report confident detections for the two largest peaks in the mass distribution but only modest evidence for the dip at $\sim$14 $M_\odot$.

4. Discussion

Previous analyses of the BBH mass spectrum by the LVK and others have found evidence for structure beyond a simple power law (Abbott et al. 2021a, 2021c). There has been considerable work exploring possible astrophysical causes of these identified features. Our aim is instead to determine, from a statistical viewpoint, whether astrophysical arguments need be invoked at all.

We first demonstrate that it is only possible for up to two of the three deviations from a power law to be explained by Poisson noise about a single power-law distribution. Therefore, at least one feature must be added on top of a power law to describe the data.

We then perform a more thorough analysis, simulating thousands of BBHs with measurement uncertainty, selection effects, and a known underlying distribution. We fit the POWER LAW + SPLINE model to the resulting catalogs and find that the data is inconsistent with a single power law, agreeing with the LVK result. However, we find that one of the previously identified features, an underdensity at $\sim 14\,M_{\odot}$, may not be present in the true astrophysical distribution. Instead, it may have been the result of a Poisson fluctuation around a simple power law, or an artifact of the models used to fit the mass spectrum. The metrics constructed in this work differ from those previously used to assess the significance of features in the mass distribution because, by virtue of comparing to several simulated catalogs, they correct for the look-elsewhere effect. This is only in mild tension with the conclusions reached by Abbott et al. (2021c), as they report "modest evidence" in favor of a dip at $14\,M_{\odot}$.

We find the other two previously identified peaks, at $\sim 10\,M_{\odot}$ and $\sim 35\,M_{\odot}$, unlikely to be the result of Poisson noise or modeling artifacts. Simulated catalogs coming from distributions that do not include these features can reproduce the height of the $\sim 10\,M_{\odot}$ peak, but not its lack of support for zero perturbation. The $\sim 35\,M_{\odot}$ peak is difficult to reproduce from featureless catalogs in any way.

Our conclusions are consistent with a recent study by Callister & Farr (2023) who fit the BBH mass distribution with an autoregressive model and find that the primary mass distribution gradually decreases as a function of mass and exhibits two local maxima with a relatively flat continuum between them. They interpret the 14 solar mass dip found by other analyses to be a flattening of the power-law index at lower masses rather than a local minimum. We also find similar results to Edelman et al. (2022b) who construct the mass distribution entirely from basis splines and find peaks at $\sim 10\,M_{\odot}$ and $\sim 35\,M_{\odot}$. The significance of the peaks near $10\,M_{\odot}$ and $35\,M_{\odot}$, as well as the lack of significance of the dip near $14\,M_{\odot}$, is also in agreement with those from Sadiq et al. (2022), Wong & Cranmer (2022).

The dip near $\sim 14\,M_{\odot}$ may be a large Poisson fluctuation or an artifact of the models used to characterize it. If it is in fact a feature of the underlying distribution, it is difficult to resolve with current observations.

The peak near $\sim 10\,M_{\odot}$ is likely an imprint of the true astrophysical distribution. Its amplitude is slightly larger than what featureless catalogs can produce with random fluctuations, and it is inconsistent with the power law that describes the rest of the BBH mass distribution at a level that only a small fraction of featureless catalogs can achieve. We therefore report moderate evidence that additional structure beyond a power law is needed to explain the peak at $\sim 10\,M_{\odot}$.

The $\sim 10\,M_{\odot}$ feature may be either an additional peak that is distinct from the one created by the underlying smoothed power law at $\sim 7\,M_{\odot}$ (Abbott et al. 2021c; Edelman et al. 2022; Tiwari 2022) or the sole peak in the region between $\sim 5\,M_{\odot}$ and $\sim 20\,M_{\odot}$ (Edelman et al. 2022b). These two possibilities can be seen in Figure 1. The former scenario is the case where we interpret the first two peaks in the orange band as distinct from one another, therefore treating the global maximum inferred by POWER LAW + SPLINE as a different feature from the global maximum inferred by POWER LAW + PEAK. In the latter scenario, the role of the perturbation function is to shift the global maximum from the value inferred by the power-law component to a slightly higher value without removing the mass distribution's support for 5–10 $M_{\odot}$ objects. A simple smoothed power law, such as that employed by the POWER LAW + PEAK model (see gray and blue bands in Figure 1), may not be flexible enough to place a global maximum at $\sim 10\,M_{\odot}$ while also fitting the correct slope at larger masses and fitting the correct merger rate below $\sim 10\,M_{\odot}$, so it places its global maximum at $\sim 7\,M_{\odot}$. This scenario, in which there is a single local maximum below $\sim 12\,M_{\odot}$, is consistent with Edelman et al. (2022b), Callister & Farr (2023), both of whom find only one significant maximum between approximately $3\,M_{\odot}$ and $12\,M_{\odot}$ using fully nonparametric methods. If this interpretation is correct and the global maximum of the BBH mass distribution is indeed offset from the minimum mass by $\sim 5\,M_{\odot}$, the upper edge of the lower mass gap may not be as morphologically simple as previously assumed (e.g., Fishbach et al. 2020a; Ezquiaga & Holz 2022; Farah et al. 2022a), making it potentially difficult to resolve with parametric models alone. The marginal evidence for the significance of the $\sim 10\,M_{\odot}$ peak is likely driven by the fact that the current observations are insufficient to distinguish between these two scenarios, which is unsurprising considering the lower sensitivity of GW detectors to low-mass events relative to high-mass events.

A peak anywhere between $\sim 7\,M_{\odot}$ and $\sim 10\,M_{\odot}$ could be indicative of particular evolutionary processes that are dominant within formation environments. van Son et al. (2022b) showed that a global maximum near this value is consistent with and robustly predicted by the stable mass transfer channel in isolated binary evolution, as stability during mass transfer requires mass ratios between the donor star and accreting compact object to be relatively symmetric, and the stellar companions to $\sim 10\ M_{\odot}$ BHs must be near this mass to form compact objects above the minimum BH mass. This may be an indication that the stable mass transfer channel operates more efficiently than the traditional common envelope channel for generating merging BBHs. If the stable mass transfer channel is indeed the cause of the global maximum in the primary mass distribution, the exact location of this global maximum will constrain the core mass fraction, mass transfer stability, and mass transfer efficiency of this process (van Son et al. 2022b). Although the dynamical formation channels with low escape velocities, such as globular clusters, struggle to produce a global maximum at $10\ M_{\odot}$ (Antonini et al. 2023), the dynamical environments with higher escape velocities may more readily produce merging BBHs with lower masses around $10\ M_{\odot}$ due to the more prevalent lower-mass BHs preferentially remaining bound to these clusters following supernova kicks.

We find that the peak centered on $35\,M_{\odot}$ is the most likely to be a feature of the true underlying distribution. This bodes well for the "spectral siren" (Farr et al. 2019; Ezquiaga & Holz 2022) method of estimating cosmological parameters from GW observations, as this peak happens to be the most informative feature for this method since it is a well-measured, somewhat-sharp feature in the mass distribution (Abbott et al. 2021d). The astrophysical process that gives rise to this feature is still a topic of discussion. The key reason for including a flexible bump-like feature in the phenomenology of parametric models, such as the POWER LAW + PEAK model used by the LVK

(Talbot & Thrane 2018), was to accommodate a potential build-up of BHs with masses just below the pair instability mass gap, as pulsational pair instability supernovae are predicted to efficiently shed material from high-mass stars with cores in the mass range of $M_{\rm core} \sim 45$–65 (Woosley 2017, 2019; Marchant et al. 2019; Renzo et al. 2020). It is difficult to reconcile the locations of the local maxima found in the BBH primary mass distribution with predictions of the pair instability process in the cores of massive stars. The largest uncertainty determining the location of the lower edge of the pair instability mass gap is the ^{12}C(α, γ)^{16}O reaction rate, which determines the abundance of oxygen in stellar cores (e.g., Farmer et al. 2019). Higher ^{12}C(α, γ)^{16}O reaction rates lead to a higher oxygen abundance in the stellar core, which will ignite explosively during core collapse and lead to (pulsational) pair instability supernovae occurring at lower core masses. However, even at 3σ deviations above the median measured value of the ^{12}C(α, γ)^{16}O reaction rate, the lower end of the mass gap only reaches $\approx$38 $M_\odot$ (Farmer et al. 2020). This is above where the measured overdensity in the observed mass spectrum occurs. This may be an indication that the peak at 35 $M_\odot$ is the result of certain isolated binary evolution scenarios (e.g., chemically homogeneous evolution; see du Buisson et al. 2020; Zevin et al. 2021; Bavera et al. 2022), another BBH formation channel entirely (e.g., globular clusters; see Antonini et al. 2023), or that stellar evolution models are missing particular ingredients that can shift the location of the pair instability gap (relaxing the assumption that the exploding stars are hydrogen-free, adjustments to convective overshooting; see, e.g., Iorio et al. 2023).

Additionally, several studies have suggested that the observed peaks in the BBH mass distribution can be explained by successive generations of hierarchical mergers (Tiwari & Fairhurst 2021; Mahapatra et al. 2022; Tiwari 2022); though, no correlation has been detected in the spin distribution of BBHs (Biscoveanu et al. 2022a), which is also necessitated by the hierarchical merger formation scenario (Fishbach et al. 2017; Gerosa & Berti 2017; Rodriguez et al. 2019; Doctor et al. 2020, 2021; Kimball et al. 2020; Gerosa et al. 2021). Additionally, for these peaks to correspond to hierarchical mergers of the same population, the dominant hierarchical pairing would have to be the first generation BH with a third generation BH (Mahapatra et al. 2022; Tiwari 2022), whereas the dominant pairing predicted by Rodriguez et al. (2019) is a first generation BH generation with a second generation BH. While it is certainly possible that GWTC-3 contains hierarchical mergers (e.g., Abbott et al. 2020b; though, also see Fishbach & Holz 2020b), the relative fraction of events formed this way is likely too small to form the structure observed in the primary mass distribution (Kimball et al. 2021), and some fine-tuning may be needed to avoid a cluster catastrophe (Zevin & Holz 2022). The exact physical reason for the overdensity at 35 $M_\odot$ therefore remains unclear. However, we confirm that it is a robust signature in the observational data; future observing runs will help to constrain its precise location, width, and possible redshift evolution.

Acknowledgments

The authors gratefully acknowledge Reed Essick for helpful insights on injections and model systematics, as well as Thomas Callister and Thomas Dent for useful comments on the manuscript. We also thank the anonymous reviewer for suggesting several helpful checks of the analysis presented here. A.M.F. is supported by the National Science Foundation Graduate Research Fellowship Program under grant No. DGE-1746045. B.E and B.F are supported in part by the National Science Foundation under grant PHY-2146528. M.Z. is supported by NASA through the NASA Hubble Fellowship grant No. HST-HF2-51474.001-A awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS5-26555. J.M.E. is supported by the European Unions Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 847523 INTERACTIONS, by VILLUM FONDEN (grant No. 37766), by the Danish Research Foundation, and under the European Unions H2020 European Research Council (ERC) Advanced Grant "Black holes: gravitational engines of discovery" grant agreement No. Gravitas101052587. D.E.H is supported by NSF grant Nos. AST-2006645 and PHY-2110507, as well as by the Kavli Institute for Cosmological Physics through an endowment from the Kavli Foundation and its founder Fred Kavli. D.E.H also gratefully acknowledges the Marion and Stuart Rice Award. This material is based upon work supported by NSF LIGO Laboratory, which is a major facility fully funded by the National Science Foundation. This research has made use of data, software and/or web tools obtained from the Gravitational Wave Open Science Center (https://www.gw-openscience.org/), a service of LIGO Laboratory, the LIGO Scientific Collaboration, and the Virgo Collaboration. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants PHY-0757058 and PHY-0823459. This work benefited from access to the University of Oregon high performance computer, Talapas.

Software: `gwpopulation` (Talbot et al. 2019), `bilby` (Ashton et al. 2019; Romero-Shaw et al. 2020), `dynesty` (Speagle 2020), `numpy` (Harris et al. 2020), `xarray` (Hoyer & Hamman 2017), `matplotlib` (Hunter 2007), `pandas` (pandas development team 2020).

Appendix A Generation of Mock Observations in GWMockCat

We describe the process used to simulate gravitational-wave event posteriors in mass and redshift, based on the procedure developed in Fishbach et al. (2020b).

This process neglects the generation of spin posteriors as this work only seeks to understand the significance of features in the mass distribution, and individual-event likelihoods are approximately separable in spin and primary mass for BBHs, and we do not model any spin populations in this work. However, spin and mass parameters are not totally uncorrelated for low-mass or high mass ratio events, so future work attempting to validate features seen in the mass ratio distribution, NSBH or BNS populations should consider simulating spin parameters as well. A lightweight, publicly available `python` package that can reproduce these mock posteriors and generate similar catalogs from arbitrary underlying populations and detector sensitivities is available for download and installation.[10] The package is called `GWMockCat`, and installation instructions, examples, and documentation are available in the git repository. Several packages exist to

[10] https://git.ligo.org/amanda.farah/mock-PE

draw events from BBH population models (e.g., Belczynski et al. 2008; Breivik et al. 2020; Riley et al. 2022), some of which also simulate GW detector selection effects (Karathanasis et al. 2022). GWMockCat complements these by additionally simulating event-level posteriors without the need to run full parameter estimation inference, saving significant computational time.

To create realizations of catalogs that would reasonably result from a known underlying astrophysical population, $p(m_1, m_2, z)$, we first make independent draws of the event parameters, $\{m_1, m_2, z\}$, from that population model. Each draw corresponds to a potential event in the catalog, although we draw many more potential events than we wish to keep since not all events generated from the astrophysical distribution will ultimately be detected. We then convert each event's redshift z and source-frame component masses to a detector-frame (redshifted) chirp mass, $\mathcal{M}_{\rm det}$, and symmetric mass ratio, η. The symmetric mass ratio and source-frame chirp mass $\mathcal{M}$ are related to the source-frame component masses via

$$\eta = \frac{m_1 m_2}{(m_1 + m_2)^2}; \tag{A1}$$

$$\mathcal{M} = \frac{(m_1 m_2)^{3/5}}{(m_1 + m_2)^{1/5}}. \tag{A2}$$

All detector-frame masses are related to their source-frame values via $M_{\rm det} = M(1 + z)$, where M can describe any parameter with units of mass (e.g., m_1, m_2, or $\mathcal{M}$).

We then utilize the basic procedure outlined in Fishbach et al. (2018), Fishbach et al. (2020b) to assign "observed" parameters for each event, using a measurement uncertainty that is correlated across parameters and a mock parameter estimation likelihood. We first calculate an optimally oriented signal-to-noise ratio (S/N) $\rho_{\rm opt}$ from the true event parameters using a characteristic power spectral density (PSD) of the LIGO Livingston detector in O3 (Abbott et al. 2020a). $\rho_{\rm opt}$ is the S/N that an event would have if it were "optimally oriented" with respect to the detector, that is, directly overhead and with its angular momentum vector pointed along the line of sight (Chen et al. 2021). In reality, GW sources have varying sky positions and angular momentum vectors. The effect on the S/N of a source's deviation from the optimal orientation can be summarized by a multiplicative constant, Θ, such that

$$\rho = \rho_{\rm opt}\Theta, \tag{A3}$$

where Θ is between zero and unity.

GW sources are typically assumed to be distributed isotropically in sky position and orientation. For a single detector, this yields a corresponding distribution for Θ, described in Finn & Chernoff (1993). Therefore, for each event i, we assign a true value $\hat{\Theta}_i$ drawn from this distribution and use it to calculate the event's true single-detector S/N $\hat{\rho}$. The set of *true* parameters for each potential event in the catalog is then $\hat{\theta}_i = \{\hat{\mathcal{M}}_{\rm det}, \hat{\eta}, \hat{\rho}, \hat{\Theta}\}_i$.

Given the true parameters, the basic procedure of generating samples from the posterior distribution of each event is to draw an observation from each event's likelihood, use that observation as the central value of the posterior distribution, and then to draw samples from that posterior, assuming a prior.

To obtain the *observed* parameters, $\theta_i^{\rm obs}$, we need the likelihood, $\mathcal{L}_{\rm total}(\theta_i^{\rm obs}|\hat{\theta}_i)$. We model each event's likelihood as

$$\mathcal{L}_{\rm total}(\theta_i^{\rm obs}|\hat{\theta}_i) = \mathcal{L}_{\mathcal{M}}(\mathcal{M}_{{\rm det},i}^{\rm obs})\mathcal{L}_{\eta}(\eta_i^{\rm obs})\mathcal{L}_{\Theta}(\Theta_i^{\rm obs})\mathcal{L}_{\rho}(\rho_i^{\rm obs}), \tag{A4}$$

where

$$\begin{aligned}
\mathcal{L}_{\mathcal{M}}(\ln(\mathcal{M}_{{\rm det},i}^{\rm obs})|\ln(\hat{\mathcal{M}}_{{\rm det},i}), \rho_i^{\rm obs}) &= \mathcal{N}(\ln(\mathcal{M}_{{\rm det},i}^{\rm obs})|\mu \\
&= \ln(\hat{\mathcal{M}}_{{\rm det},i}), \sigma = \sigma_i^{\mathcal{M}}(\rho_i^{\rm obs})),\\
\mathcal{L}_{\eta}(\eta_i^{\rm obs}|\hat{\eta}_i, \rho_i^{\rm obs}) &= \mathcal{N}(\eta_i^{\rm obs}|\mu = \hat{\eta}_i, \sigma = \sigma_i^{\eta}, (\rho_i^{\rm obs})),\\
\mathcal{L}_{\Theta}(\Theta_{{\rm obs},i}|\hat{\Theta}_i, \rho_i^{\rm obs}) &= \mathcal{N}(\Theta_i^{\rm obs}|\mu = \hat{\Theta}_i, \sigma = \sigma_i^{\Theta}(\rho_i^{\rm obs})), \text{ and}\\
\mathcal{L}_{\rho}(\rho_i^{\rm obs}|\hat{\rho}_i) &= \mathcal{N}(\rho_i^{\rm obs}|\mu = \hat{\rho}_i, \sigma = \sigma_i^{\rho}).
\end{aligned} \tag{A5}$$

Here, $\mathcal{N}(\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ.

The standard deviations are determined by assuming the uncertainties on all parameters except for the S/N scale inversely with $\rho_{\rm obs}$ (Veitch et al. 2015). In stationary, Gaussian noise, we expect the matched-filter S/N in a single detector to have unit variance (Allen et al. 2012), i.e., $\sigma_i^{\rho} = 1$ for all i. We therefore draw $\rho_{\rm obs}$ for each event from $\mathcal{L}^{\rho}(\rho_i^{\rm obs}|\rho_i)$. This observed S/N will serve as the detection statistic that determines whether each event is observable. We assume events that pass an S/N threshold of $\rho_{{\rm obs},i} > 8$ in a single detector are detected. In this way, we allow for events near the threshold to fluctuate above or below the threshold, emulating the actual noise process in the detectors. Of the events that make it through detection, we randomly select 69 of them to constitute a mock catalog with the same number of BBHs as were analyzed by Abbott et al. (2021c). The standard deviations for $\mathcal{M}_{\rm det}$, η, and Θ of the detected events are calculated via

$$\begin{aligned}
\sigma_i^{\mathcal{M}}(\rho_i^{\rm obs}) &= u_{\mathcal{M}}/\rho_i^{\rm obs},\\
\sigma_i^{\eta}(\rho_i^{\rm obs}) &= u_{\eta}/\rho_i^{\rm obs},\\
\sigma_i^{\Theta}(\rho_i^{\rm obs}) &= u_{\Theta}/\rho_i^{\rm obs},
\end{aligned} \tag{A6}$$

where we have chosen $u_{\mathcal{M}} = 0.08\,M_\odot$, $u_\eta = 0.022$, and $u_\Theta = 0.21$ to match uncertainties in these parameters typical of events observed in O3.

Observed values for all parameters are drawn from Equation (A4) with standard deviations defined in Equation (A6). With $\theta_i^{\rm obs}$ in hand, we are now ready to construct a posterior distribution. We apply the following priors:

$$\begin{aligned}
\pi(\mathcal{M}_{\rm det}) &= U(0\,M_\odot, 500\,M_\odot),\\
\pi(\eta) &= U(0, 0.25),\\
\pi(\Theta) &= U(0, 1),\\
\text{and } \pi(\rho) &= U(0, 300),
\end{aligned} \tag{A7}$$

where $U(x_1, x_2)$ is the uniform distribution with lower bound x_1 and upper bound x_2. The bounds on η and Θ are chosen because those parameters are only physically defined in the domains [0, 0.25] and [0, 1], respectively. Neither $\mathcal{M}$ nor ρ are defined below zero, but the upper bounds were chosen somewhat arbitrarily: they must only be large enough that the likelihood has minimal support above them. The posterior distributions for each parameter are then Gaussians centered on the observed value, with standard deviations defined in Equation (A6). They are therefore the same as the distributions in Equation (A5), but with the role of the true and observed values switched. We then simulate multiple-dimensional posterior samples for each event by drawing 5000 independent samples[11] of detector-frame chirp mass, symmetric mass ratio, and Θ from the posterior. Explicitly,

$$\begin{aligned} \ln \mathcal{M}_{\mathrm{det},i} &\sim \mathcal{N}(\mu = \ln(\mathcal{M}^{\mathrm{obs}}_{\mathrm{det},i}),\ \sigma = \sigma_i^{\mathcal{M}}); \\ \eta_i &\sim \mathcal{N}(\mu = \eta_i^{\mathrm{obs}},\ \sigma = \sigma_i^{\eta}); \\ \Theta_i &\sim \mathcal{N}(\mu = \Theta_i^{\mathrm{obs}},\ \sigma = \sigma_i^{\Theta}); \\ \rho_i &\sim \mathcal{N}(\mu = \rho_i^{\mathrm{obs}},\ \sigma = \sigma_i^{\rho}). \end{aligned} \tag{A8}$$

Realistic correlations between other parameters such as component masses and redshift are obtained by transforming samples in $\{\mathcal{M}_{\mathrm{det}}, \eta, \Theta, \rho\}$–space to $\{m_1, m_2, z\}$–space. When necessary, we convert between luminosity distance and redshift using the cosmological parameters presented in Planck Collaboration et al. (2016) so as to maintain consistency with the conventions used in Abbott et al. (2019b, 2021d, 2021b).

The induced prior on m_1, m_2, and z is therefore not uniform in those parameters. This is reasonable, so long as users appropriately transform the prior when doing population inference on source-frame component masses and redshift. We therefore provide a module in `GWMockCat` that performs these transformations. For the case of this analysis, we opt to reweigh the samples to a prior that is uniform in detector-frame component mass and proportional to the square of the luminosity distance in order to mimic the priors used in the standard LVK analysis (Abbott et al. 2019b, 2021d, 2021b).

The fact that Equation (A4) is separable up to dependence on $\rho_{\mathrm{obs},i}$ means that once $\rho_{\mathrm{obs},i}$ is calculated for a given event, samples for $\mathcal{M}_{\mathrm{det}}$, η_{obs}, Θ_{obs}, and ρ_{obs} can be drawn independently from each other. This approximate independence is due, in part, to the fact that detector-frame chirp mass, symmetric mass ratio, S/N, and Θ are the best-measured parameters of any compact binary coalescence signal. This fact saves considerable computational resources, allowing for many mock event posteriors to be generated quickly on a single CPU.[12]

We generate sensitivity estimates along with our mock catalogs to ensure that the selection function is calculated consistently to the event selection criteria (Essick & Fishbach 2022). To do this, we draw 2×10^7 independent samples in m_1, m_2, z, and Θ from the following distribution:

$$p(m_1, m_2, z, \Theta) \propto m_1^{\alpha} \left(\frac{m_2}{m_1}\right)^{\beta} \frac{dV_c}{dz} (1 + z)^{\kappa - 1} p(\Theta), \tag{A9}$$

where we have chosen $\alpha = 2.35$, $\beta = 1.70$, and $\kappa = 2.7$, and $p(\Theta)$ is the distribution described in Finn & Chernoff (1993), which corresponds to isotropically oriented sources that are also isotropically positioned on the sky. We truncate this distribution below $m_2 = 1\,M_{\odot}$, above $m_1 = 200\,M_{\odot}$, and above $z = 4$, and confirm that there are no mock posterior samples outside of those ranges. We will refer to these draws as *injections*. We then calculate an optimally oriented S/N for each injection using the same PSD as was used for the mock observations, and compute a true S/N using Equation (A3). We emulate noise fluctuations in S/N in the same way we do for mock observations, namely by using Equation (A5), so that each injection has a corresponding observed S/N. Injections can then be subject to the same selection criteria as our mock observations when performing a population inference (in our case, $\rho_{\mathrm{obs}} > 8$).

We validate this process by constructing a mock catalog from a known distribution with fixed hyperparameters, and then fitting the same distribution to our mock catalog, but allowing the hyperparameters to vary. We then verify that the recovered hyperparameters are consistent with those used to generate the mock catalog. The result is shown in Appendix B, along with additional validation studies.

Appendix B
Validation of Mock Catalogs

In this appendix, we validate the process of creating mock event posteriors and catalogs from a known underlying population outlined in Appendix A. For this process, we use the same simulated catalogs utilized in Section 3.3. The simulated underlying population is described by $p_{\mathrm{mock}}(m_1, m_2, z|\Lambda_{\mathrm{mock}})$, where Λ_{mock} is the set of hyperparameters $\{\alpha, \delta, m_{\min}, m_{\max}, \beta, \kappa\}$,

$$\begin{aligned} p_{\mathrm{mock}}(m_1, m_2, z|\Lambda_{\mathrm{mock}}) &\propto p(m_1|\alpha, \delta, m_{\min}, m_{\max}) \\ &\times p(m_2|m_1, \beta) p(z|\kappa), \end{aligned} \tag{B1}$$

and the individual mass and redshift distributions are given by the following:

$$p(m_1|\alpha, \delta, m_{\min}, m_{\max}) \propto \begin{cases} 0 & \text{if } m < m_{\min} \\ m_1^{-\alpha} \frac{1}{1 + f(m - m_{\min}, \delta)} & \text{if } m_{\min} \leqslant m < m_{\min} + \delta \\ m_1^{-\alpha} & \text{if } m \geqslant m_{\min} + \delta \\ 0 & \text{if } m > m_{\max} \end{cases}, \tag{B2}$$

$$p(m_2|m_1, \beta) \propto \left(\frac{m_2}{m_1}\right)^{\beta}, \tag{B3}$$

Table 1
Hyperparameter Values for the Underlying Population of Mock Catalogs Described by SMOOTHED POWER LAW (Equations (B1)–B4)

Parameter	Description	Value
β	Spectral index for the power law of the mass ratio distribution.	1.70
α	Negative spectral index for the power law of the primary mass distribution.	3.14
$m_{\min}$	Minimum mass of the primary mass distribution.	$4.56\,M_{\odot}$
$m_{\max}$	Maximum mass of the primary mass distribution.	$81.08\,M_{\odot}$
δ	Range of mass tapering at the lower end of the mass distribution.	$5.96\,M_{\odot}$
κ	Spectral index for the power-law factor of the redshift distribution.	2.7

[11] We use 5000 samples to optimize the speed of population inference while also ensuring the number of effective samples used for Monte Carlo sums in the population inference always satisfies the criterion outlined in Farr (2019). That criterion has since been shown to be insufficient and has been superseded by Essick & Farr (2022), but we utilize the former for consistency with the analysis performed in Abbott et al. (2021c). However, users of the `GWMockCat` package can easily modify the number of posterior samples to suit their needs.

[12] For example, a catalog of 100 events can be generated in $\mathcal{O}(10)$ s.

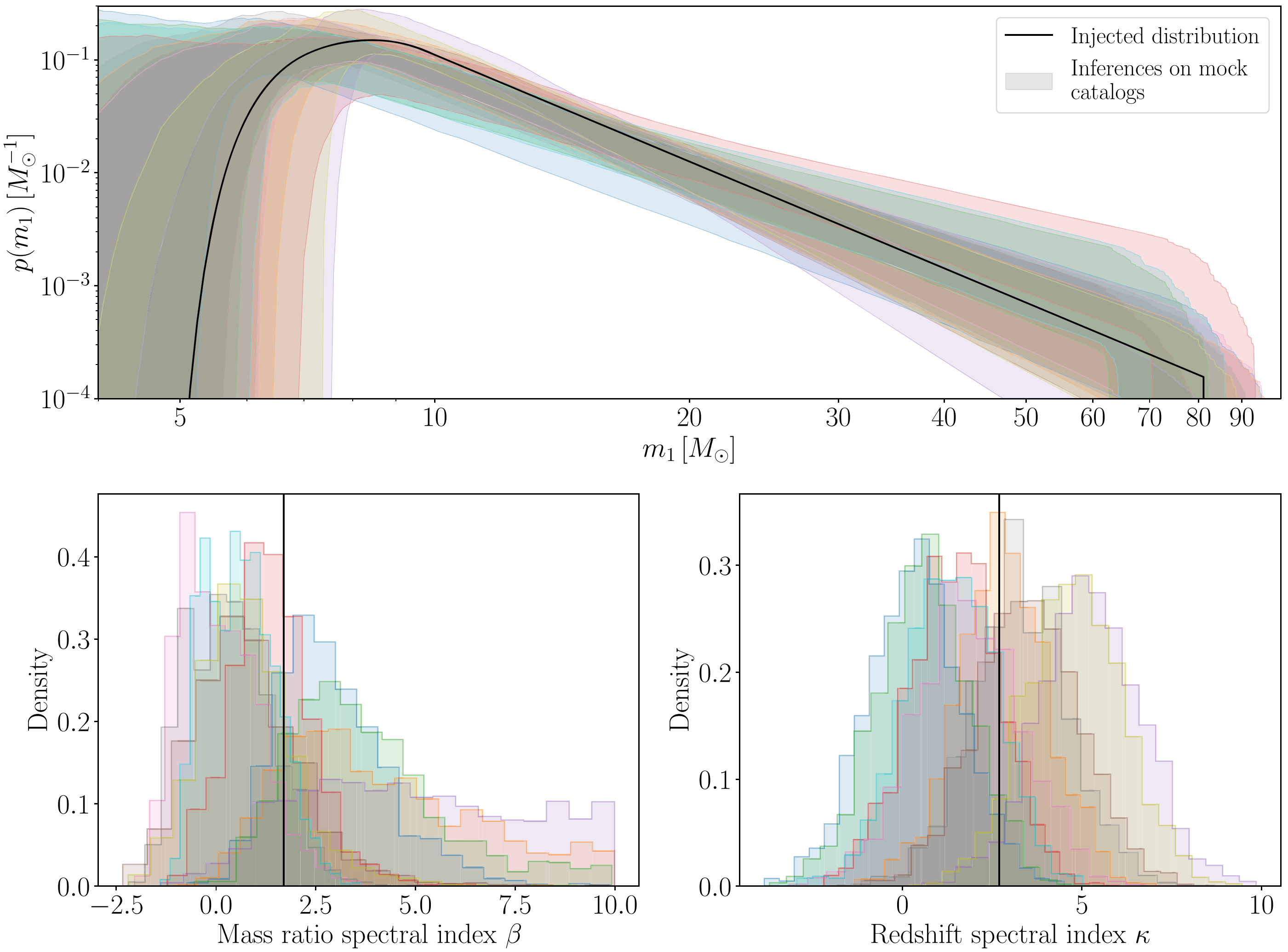

Figure 6. Injected (solid black line) and recovered (colored shaded bands) distributions for 10 mock catalogs. Top: probability density function of primary masses. Bottom left: hyperposterior distribution for β, the power-law spectral index of the mass ratio distribution. Bottom right: hyperposterior distribution for κ, the spectral index of the power-law factor in the redshift distribution.

and

$$p(z|\kappa) \propto \begin{cases} 0 & \text{if } (z < 0) \cup (z > z_{\max}) \\ \frac{dV_c}{dz}(1+z)^{\kappa-1} & \text{otherwise} \end{cases}. \tag{B4}$$

This is equivalent to the POWER LAW + PEAK model in Abbott et al. (2021c), Abbott et al. (2021a), with λ_{peak} set to 0. We will call the population model described by Equations (B1)–(B4) SMOOTHED POWER LAW. We generate catalogs from the model that results from setting Λ_{mock} to the values provided in Table 1. These values were chosen by fitting this population model to GWTC-3 (gray band in Figure 1) and obtaining the median a posteriori value for each hyperparameter.

We validate the mock catalogs' generation by fitting them with SMOOTHED POWER LAW and allowing the hyperparameters to be inferred from the mock data. We then determine whether the inferred values of the hyperparameters are consistent with the values in Table 1. We fit 100 mock catalogs of 69 events each, 10 results of which are shown in Figure 6. While there is noticeable scatter about the injected value, it is generally consistent with the recovered mass distributions: the hyperparameters of the underlying mass distribution fall within the inferred mass hyperparameters' 90% credible intervals 89.6% of the time. We therefore conclude that any biases that the mock posterior generation process introduces in the mass distribution are subdominant to the statistical uncertainties of the fit.

To further explore systematic differences caused by mock catalog generation that may be subdominant to the considerable statistical uncertainties resulting from a fit to only 69 events, we fit SMOOTHED POWER LAW to a single catalog of that is 5 times larger. The result is shown in Figure 7. The hyperparameters of the underlying distribution seem to be consistent with the inferred hyperposterior, so we conclude that our mock event posterior generation process produces biases subdominant to measurement uncertainty typical of 345-event catalogs. We therefore find this method of generating mock catalogs sufficient to test the significance of features identified in the mass distribution of GWTC-3.

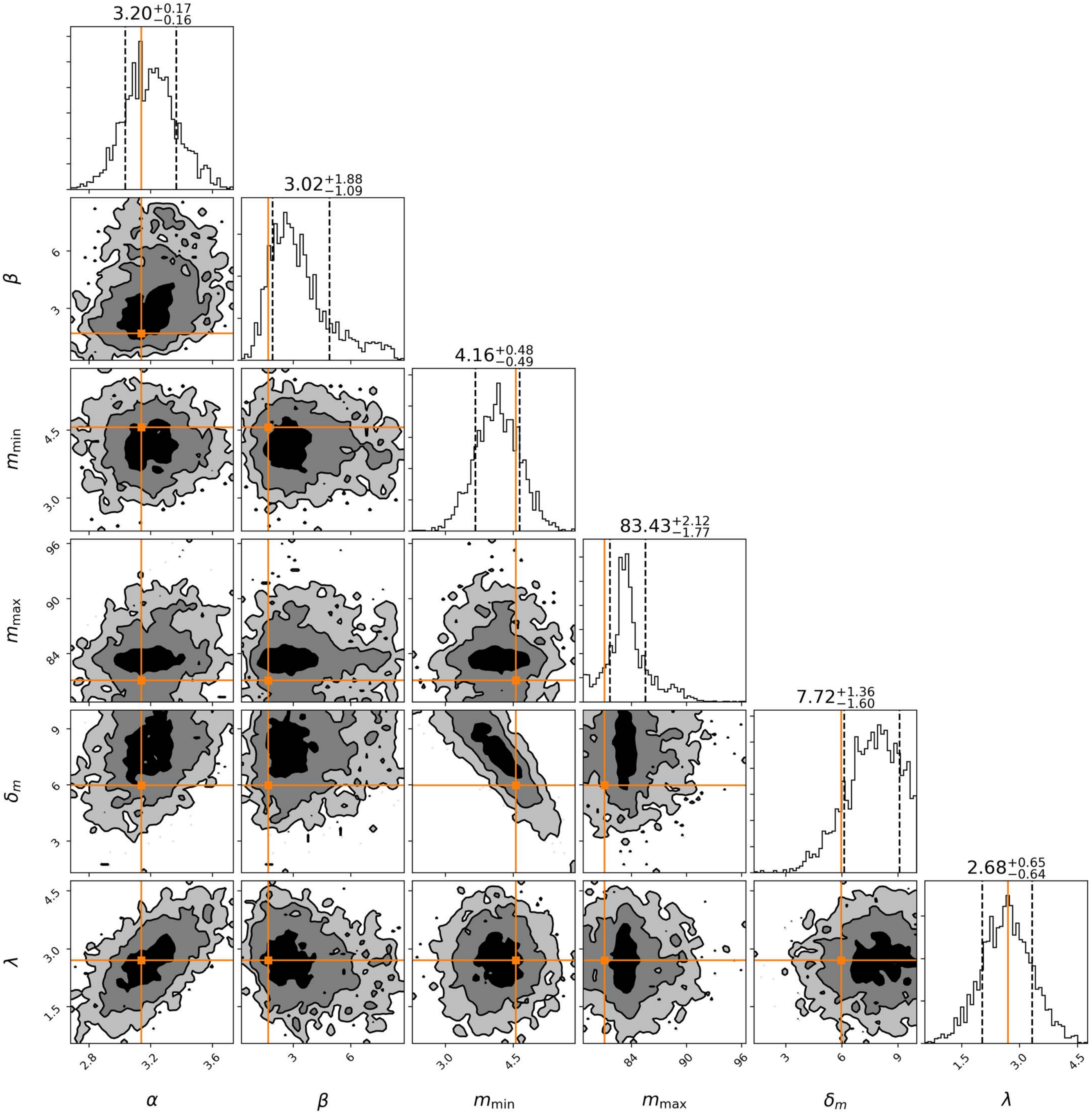

Figure 7. A corner plot of the inferred hyperposterior from a fit to a mock catalog with 345 events. The injected values are shown in orange. The recovered hyperposterior is consistent with the injected population.

Appendix C
Accuracy of Mock Catalogs When Used in a Population Analysis

As a second check of `GWMockCat`'s ability to simulate catalogs accurately enough to be used in a population analysis of the mass distribution of BBHs, we recreate GWTC-3 with `GWMockCat`. We then compare POWER LAW + SPLINE's fit to this mock catalog with its fit to the posterior samples released by the LVK for GWTC-3 (LIGO Scientific Collaboration et al. 2021b). We find that the two resulting mass distributions are consistent, and therefore conclude that the approximate prescriptions used in `GWMockCat` are sufficient to probe the mass distribution of BBHs, at least for current GW detector sensitivities.

To simulate GWTC-3, we reweight all GWTC-3 events to the same prior as used for sampling in `GWMockCat`, namely uniform in detector-frame chirp mass, uniform in symmetric mass ratio, and uniform in sky angle, as defined in Equation (A7). We then take the mean of the reweighted detector frame chirp mass, symmetric mass ratio, sky angle, and single-detector S/N posteriors as the observed parameters $\theta_i^{\rm obs}$

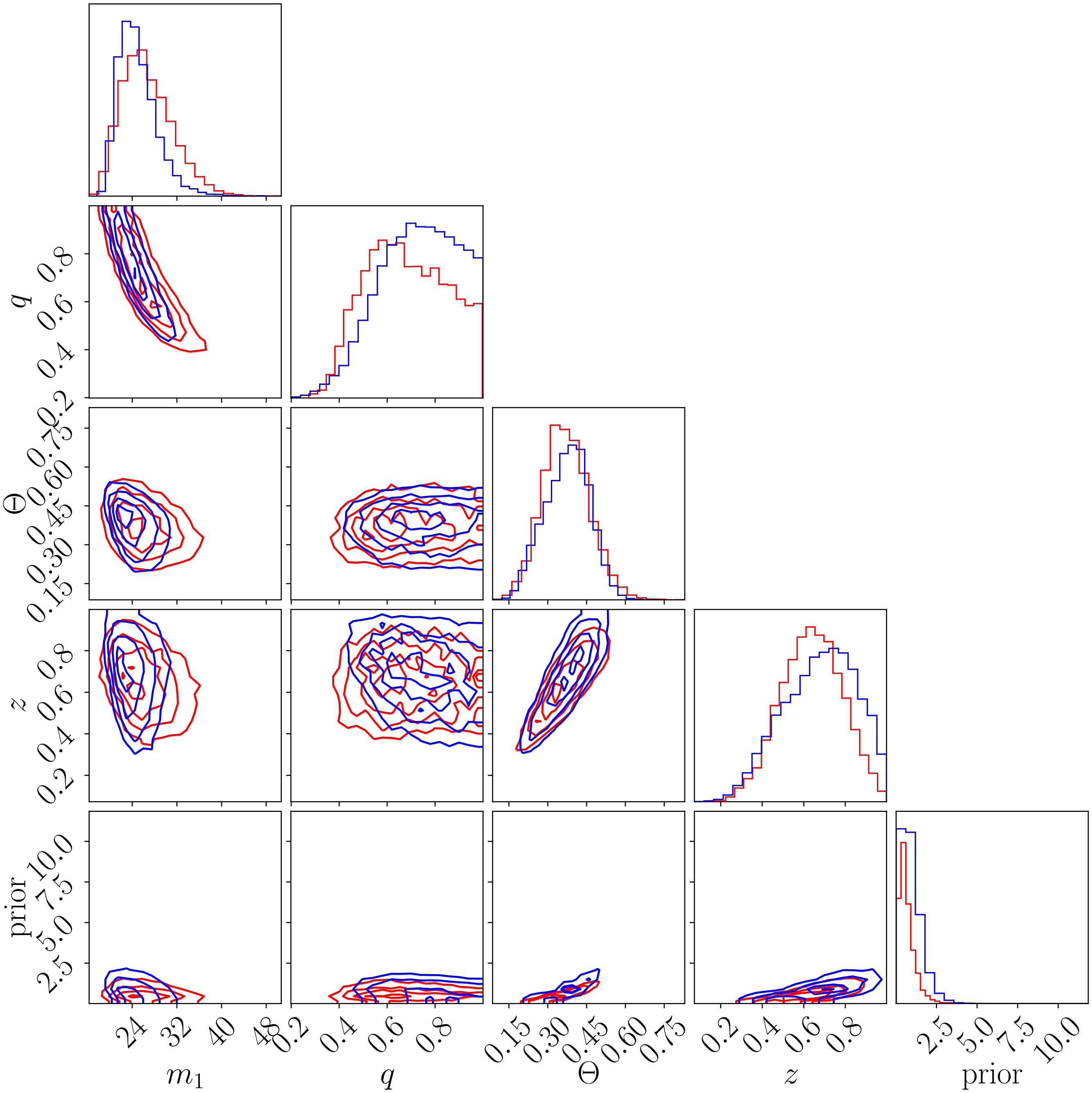

Figure 8. Mock posteriors simulated with `GWMockCat` (red) compared to posteriors made with full parameter estimation as released by the LVK (blue) for the event GW191215_223052. The two sets of posterior samples appear consistent to the level needed for a population analysis.

for each event. Using the priors defined in Equation (A7) and likelihoods defined in Equations (A5) and (A6), we construct mock posterior distributions on detector frame chirp mass, symmetric mass ratio, sky angle, and single-detector S/N. We then convert these parameters to source-frame component masses and redshift. Finally, we reweight back to the standard prior used in LVK's parameter estimation process. The end result of this is shown in Figure 8 for GW191215_223052, an event that was chosen at random from the catalog and happened to have a primary mass and redshift near the mode of the detected events. The mock posteriors appear consistent with the true posteriors, and the degeneracies between parameters seem to be suitably captured. This behavior is qualitatively similar for all simulated events; though, some had mock posteriors that were slightly more consistent with the full parameter estimation posterior samples, and some had mock posteriors that were slightly less consistent.

Any population analysis must define criteria for inclusion in the population. We apply two different criteria and report the results of both in Figure 9. The first, a cut at $S/N > 8$ in Livingston, is chosen to be analogous to the detection criteria of the other mock catalogs, which use a single-detector S/N cut since it is impractical to run all of the pipelines necessary to produce a FAR on mock data. The second criterion, a FAR cut at 1 yr^{-1} was used to be consistent with the analysis done by the LVK on GWTC-3 (Abbott et al. 2021c). Either choice is reasonable because the selection function is known with respect to both S/N and FAR. We can therefore use the publicly

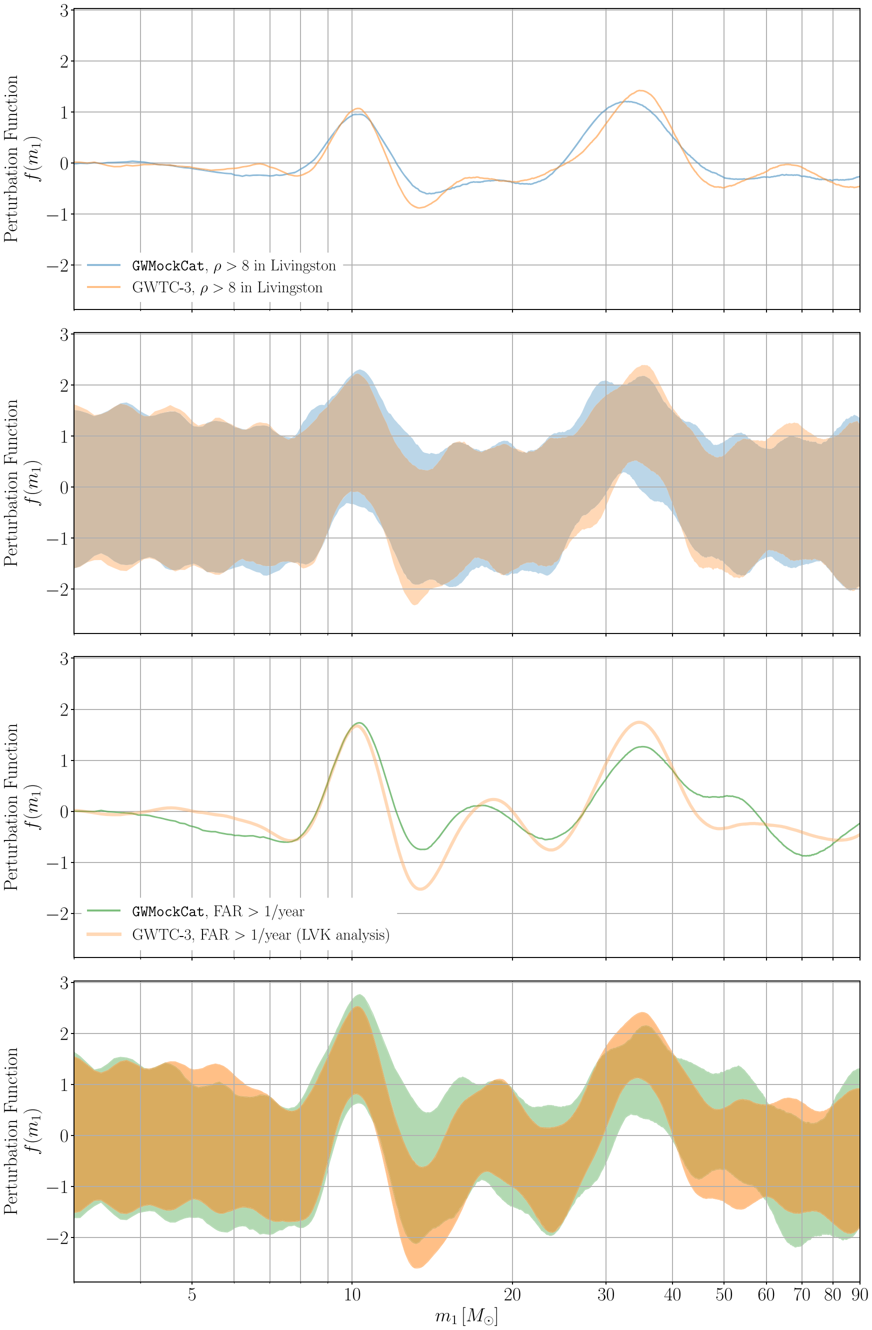

Figure 9. Perturbation functions of a population fit to GWTC-3 when using LVK-released posterior samples for each event (orange), and when using posterior samples simulated by GWMockCat (blue and green). Shaded bands correspond to 90% credible intervals, and solid lines are the medians. The two sets of figures correspond to two selection criteria, one which is analogous to the one used for mock catalogs in this work (S/N >8, top two panels), and one which is identical to that used for the LVK analysis (FAR<1yr^{-1}, bottom two panels). The main difference between the selection criteria is that they result in catalog sizes that are different by nearly a factor of 2, and therefore, the statistical error on the perturbation function is noticeably different between them. Using the same number of events in the simulated catalog as the LVK-released catalog results in a perturbation function that is similar in amplitude to the LVK-released population analysis. We conclude that the prescriptions used in GWMockCat are sufficient for the purposes of population analyses. We use the green curves in the bottom two panels for the analysis presented in the body of this paper, and the orange curves in the bottom two panels for the analysis presented in this appendix.

released sensitivity estimates (LIGO Scientific Collaboration et al. 2021a) to reconstruct the underlying, or astrophysical distribution of BBHs from either of these catalogs. All of the metrics of feature significance presented in Section 3 make use of this astrophysical distribution when comparing GWTC-3's population fit to that of mock catalogs.

The first criterion (single-detector S/N cut) produced a final catalog with many fewer events than the second criterion (FAR cut), with the former resulting in 36 events and the latter resulting in 69 events. Therefore, the population analysis performed on the catalog selected by S/N has much wider hyperposteriors than the one performed on the catalog selected by FAR. However, *these two catalogs do not appear to be systematically biased with respect to one another, nor are they systematically biased with respect to the LVK-released analysis*. This is again because the selection function is known with respect to both of these criteria, and therefore, the reconstructed astrophysical distributions are consistent.

Interestingly, the mock catalog with only 36 events in it still finds the peak at $\sim 35\,M_\odot$ to be significant, with the perturbation function excluding zero to <5%. However, other features are not well enough resolved to appear significant with only 36 events. We find very little difference between the full parameter estimation perturbation function fit (orange bands in Figure 9) and the mock catalog (blue band) in the case of a single-detector S/N cut (top two panels). We take this to mean that, for high-S/N events, the mock parameter estimation is sufficient for use in population analyses of the mass distribution, at least for O3-like detector sensitivities.

Using the same events in the simulated catalog as the LVK-released catalog results in a perturbation function that is similar in amplitude to the LVK-released analysis (lower two panels of Figure 9). However, the width of the perturbation function's hyperposterior is slightly inflated in the mock catalog case. This may be because the mock parameter estimation scales event posterior widths inversely with S/N, so the mock posterior widths are overestimated with respect to full parameter estimation for the events that meet the FAR threshold but have low S/N.

In order to be as consistent as possible in our comparisons of mock catalogs to GWTC-3, we use the perturbation function obtained by analyzing the `GWMockCat` version of GWTC-3 for all comparisons to mock catalogs in Section 3. In this appendix, we repeat the analysis performed in Section 3, but instead used the perturbation function released by the LVK in Abbott et al. (2021c) in lieu of the perturbation function obtained by fitting POWER LAW + SPLINE to the `GWMockCat` version of GWTC-3. In other words, the main text uses the green curves in Figure 9, and we repeat the analysis using the orange curves in the bottom two panels of Figure 9 in this appendix.

We find that using the LVK-released perturbation function increases the Bayesian significance of all features relative to the `GWMockCat` reproduction. This is consistent with the conclusions reached in Figure 9: all hyperposteriors narrow slightly when using the LVK-released version of parameter estimation, but there is no systematic shift as a function of primary mass or any other parameter. When using the LVK-released perturbation function, none of the 300 $\{g_j(f_{\max})\}$ exclude zero to the same percentile as $g_{\mathrm{GWTC-3}}(f(35\,M_\odot))$ or $g_{\mathrm{GWTC-3}}(f(10\,M_\odot))$, and 1.3% of the $\{g_j(f_{\max})\}$ exclude zero to the same percentile as $g_{\mathrm{GWTC-3}}(f(14\,M_\odot))$. These values are smaller than those presented in Section 3.3.2, but lead to the same conclusions: the $10\,M_\odot$ and $35\,M_\odot$ peaks are difficult to reproduce with featureless catalogs, but the $14\,M_\odot$ dip is not. Performing full parameter estimation on mock catalogs would also likely narrow the hyperposteriors for those catalogs, in turn increasing the significance of the peaks seen in their the perturbation functions. If this were to be the case, the fraction of mock catalogs that can reproduce features in the GWTC-3 distribution would likely be similar to those found in Section 3.3.2. A reproduction of the analysis presented in Section 3.3.1 with the LVK-released perturbation function also finds similar results to that done on the `GWMockCat` perturbation function. We therefore conclude that the results presented Section 3 are robust to the procedure used for parameter estimation of GWTC-3 events.

ORCID iDs

Amanda M. Farah https://orcid.org/0000-0002-6121-0285
Bruce Edelman https://orcid.org/0000-0001-7648-1689
Michael Zevin https://orcid.org/0000-0002-0147-0835
Maya Fishbach https://orcid.org/0000-0002-1980-5293
Jose María Ezquiaga https://orcid.org/0000-0002-7213-3211
Ben Farr https://orcid.org/0000-0002-2916-9200
Daniel E. Holz https://orcid.org/0000-0002-0175-5064

References

Aasi, J., Abbott, B. P., Abbott, R., et al. 2015, CQGra, 32, 074001
Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2019a, ApJL, 882, L24
Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2020a, LRR, 23, 3
Abbott, B. P., Abbott, R., Abbott, T. D, et al. 2019b, PhRvX, 9, 031040
Abbott, R., Abbott, T. D., Abraham, S., et al. 2020b, PhRvL, 125, 101102
Abbott, R., Abbott, T. D., Abraham, S., et al. 2021a, ApJL, 913, L7
Abbott, R., Abbott, T. D., Abraham, S., et al. 2021d, PhRvX, 11, 021053
Abbott, R., Abbott, T. D., Acernese, F., et al. 2021b, arXiv:2111.03606
Abbott, R., Abbott, T. D., Acernese, F., et al. 2023a, PhRvX, 13, 011048
Abbott, R., Abe, H., Acernese, F., et al. 2023b, ApJ, 949, 76
Abbott, R., Abe, H., Acernese, F., et al. 2021c, arXiv:2112.06861
Acernese, F., Agathos, M., Agatsuma, K., et al. 2014, CQGra, 32, 024001
Allen, B., Anderson, W. G., Brady, P. R., Brown, D. A., & Creighton, J. D. E. 2012, PhRvD, 85, 122006
Antonini, F., Gieles, M., Dosopoulou, F., & Chattopadhyay, D. 2023, MNRAS, 522, 466
Ashton, G., Hübner, M., Lasky, P. D., et al. 2019, ApJS, 241, 27
Barkat, Z., Rakavy, G., & Sack, N. 1967, PhRvL, 18, 379
Bavera, S. S., Fragos, T., Zapartas, E., et al. 2022, A&A, 657, L8
Belczynski, K., Heger, A., Gladysz, W., et al. 2016, A&A, 594, A97
Belczynski, K., Kalogera, V., Rasio, F. A., et al. 2008, ApJS, 174, 223
Biscoveanu, S., Callister, T. A., Haster, C.-J., et al. 2022a, ApJL, 932, L19
Biscoveanu, S., Landry, P., & Vitale, S. 2022b, MNRAS, 518, 5298
Breivik, K., Coughlin, S., Zevin, M., et al. 2020, ApJ, 898, 71
Callister, T. A., & Farr, W. M. 2023, arXiv:2302.07289
Chen, H.-Y., Holz, D. E., Miller, J., et al. 2021, CQGra, 38, 055010
Chernoff, D. F., & Finn, L. S. 1993, ApJL, 411, L5
Doctor, Z., Farr, B., & Holz, D. E. 2021, ApJL, 914, L18
Doctor, Z., Wysocki, D., O'Shaughnessy, R., Holz, D. E., & Farr, B. 2020, ApJ, 893, 35
du Buisson, L., Marchant, P., Podsiadlowski, P., et al. 2020, MNRAS, 499, 5941
Eddington, A. S. 1913, MNRAS, 73, 359
Edelman, B., Doctor, Z., & Farr, B. 2021, ApJL, 913, L23
Edelman, B., Doctor, Z., Godfrey, J., & Farr, B. 2022, ApJ, 924, 101
Edelman, B., Farr, B., & Doctor, Z. 2023, ApJ, 946, 14
Essick, R., & Farr, W. 2022, arXiv:2204.00461
Essick, R., & Fishbach, M. 2022, On the Consistency of Parameter Estimation and Selection Functions in Mock Catalogs LIGO-T2200210-v4, LIGO Document Control Center
Ezquiaga, J. M., & Holz, D. E. 2021, ApJL, 909, L23
Ezquiaga, J. M., & Holz, D. E. 2022, PhRvL, 129, 061102
Farah, A., Fishbach, M., Essick, R., Holz, D. E., & Galaudage, S. 2022a, ApJ, 931, 108
Farah, A. M., Edelman, B., Zevin, M., et al. 2022b, Data Release for "Things that might Go Bump in the Night: Assessing Structure in the Binary Black Hole Mass Spectrum" v1, Zenodo, doi:10.5281/zenodo.7411991
Farah, A. M., Fishbach, M., Edelman, B., Zevin, M., & Ezquiaga, J. M. 2022c, GWMockCat v1.0, Zenodo, doi:10.5281/zenodo.7570191
Farmer, R., Renzo, M., de Mink, S., Fishbach, M., & Justham, S. 2020, ApJL, 902, L36
Farmer, R., Renzo, M., de Mink, S. E., Marchant, P., & Justham, S. 2019, ApJ, 887, 53
Farr, W. M. 2019, RNAAS, 3, 66
Farr, W. M., Fishbach, M., Ye, J., & Holz, D. E. 2019, ApJ, 883, L42
Farr, W. M., Sravan, N., Cantrell, A., et al. 2011, ApJ, 741, 103
Finn, L. S., & Chernoff, D. F. 1993, PhRvD, 47, 2198
Fishbach, M., Doctor, Z., Callister, T., et al. 2021, ApJ, 912, 98
Fishbach, M., Essick, R., & Holz, D. E. 2020a, ApJL, 899, L8
Fishbach, M., Farr, W. M., & Holz, D. E. 2020b, ApJL, 891, L31
Fishbach, M., & Holz, D. E. 2017, ApJL, 851, L25
Fishbach, M., & Holz, D. E. 2020a, ApJL, 891, L27
Fishbach, M., & Holz, D. E. 2020b, ApJL, 904, L26
Fishbach, M., Holz, D. E., & Farr, B. 2017, ApJL, 840, L24
Fishbach, M., Holz, D. E., & Farr, W. M. 2018, ApJL, 863, L41
Fowler, W. A., & Hoyle, F. 1964, ApJS, 9, 201
Fryer, C. L., Belczynski, K., Wiktorowicz, G., et al. 2012, ApJ, 749, 91

Gerosa, D., & Berti, E. 2017, PhRvD, 95, 124046
Gerosa, D., Giacobbo, N., & Vecchio, A. 2021, ApJ, 915, 56
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Hartigan, J. A., & Hartigan, P. M. 1985, AnSta, 13, 70
Heger, A., Fryer, C. L., Woosley, S. E., Langer, N., & Hartmann, D. H. 2003, ApJ, 591, 288
Heger, A., & Woosley, S. E. 2002, ApJ, 567, 532
Hoyer, S., & Hamman, J. 2017, JOSS, 5, 10
Hunter, J. D. 2007, CSE, 9, 90
Iorio, G., Costa, G., Mapelli, M., et al. 2023, MNRAS, 524, 426
Karathanasis, C., Mukherjee, S., & Mastrogiovanni, S. 2023, MNRAS, 523, 4539
Karathanasis, C., Revenu, B., Mukherjee, S., & Stachurski, F. 2022, arXiv:2210.05724
Kimball, C., Talbot, C., Berry, C. P. L., et al. 2020, ApJ, 900, 177
Kimball, C., Talbot, C., Berry, C. P. L., et al. 2021, ApJL, 915, L35
Landry, P., & Read, J. S. 2021, ApJL, 921, L25
Li, A., Miao, Z., Han, S., & Zhang, B. 2021, ApJ, 913, 27
LIGO Scientific Collaboration, & Virgo Collaboration 2022, GWTC-2.1: Deep Extended Catalog of Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run—Parameter Estimation Data Release, v2, Zenodo, doi:10.5281/zenodo.6513631
LIGO Scientific Collaboration, Virgo Collaboration, & KAGRA Collaboration 2021a, GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run—O3 Search Sensitivity Estimates, v2, Zenodo, doi:10.5281/zenodo.5546676
LIGO Scientific Collaboration, Virgo Collaboration, & KAGRA Collaboration 2021b, GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run—Parameter Estimation Data Release, Zenodo, doi:10.5281/zenodo.5546663 Version 1
Loredo, T. 2009, AAS Meeting, 213, 211.04
Mahapatra, P., Gupta, A., Favata, M., Arun, K. G., & Sathyaprakash, B. S. 2022, arXiv:2209.05766
Malmquist, K. G. 1922, MeLuF, 100, 1
Malmquist, K. G. 1925, MeLuF, 106, 1
Mandel, I., Farr, W. M., Colonna, A., et al. 2017, MNRAS, 465, 3254
Mandel, I., Farr, W. M., & Gair, J. R. 2019, MNRAS, 486, 1086
Mandel, I., & Müller, B. 2020, MNRAS, 499, 3214
Marchant, P., Renzo, M., Farmer, R., et al. 2019, ApJ, 882, 36
McLachlan, G., & Peel, D. 2000, Finite Mixture Models (New York: Wiley)
Messenger, C., & Read, J. 2012, PhRvL, 108, 091101
Özel, F., Psaltis, D., Narayan, R., & McClintock, J. E. 2010, ApJ, 725, 1918
pandas development team, T 2020, pandas-dev/pandas: Pandas, v1.4.3, Zenodo, doi:10.5281/zenodo.3509134
Patton, R. A., Sukhbold, T., & Eldridge, J. J. 2022, MNRAS, 511, 903
Payne, E., & Thrane, E. 2023, PhRvR, 5, 023013
Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, A&A, 594, A13
Renzo, M., Farmer, R., Justham, S., et al. 2020, A&A, 640, A56
Riley, J., Agrawal, P., Barrett, J. W., et al. 2022, ApJS, 258, 34
Rinaldi, S., & Del Pozzo, W. 2022, MNRAS, 509, 5454
Rodriguez, C. L., Zevin, M., Amaro-Seoane, P., et al. 2019, PhRvD, 100, 043027
Romero-Shaw, I. M., Talbot, C., Biscoveanu, S., et al. 2020, MNRAS, 499, 3295
Sadiq, J., Dent, T., & Wysocki, D. 2022, PhRvD, 105, 123014
Siegel, J. C., Kiato, I., Kalogera, V., et al. 2023, ApJ, 954, 212
Speagle, J. S. 2020, MNRAS, 493, 3132
Stevenson, S., Ohme, F., & Fairhurst, S. 2015, ApJ, 810, 58
Talbot, C., Smith, R., Thrane, E., & Poole, G. B. 2019, PhRvD, 100, 043030
Talbot, C., & Thrane, E. 2018, ApJ, 856, 173
Taylor, S. R., Gair, J. R., & Mandel, I. 2012, PhRvD, 85, 023535
Thrane, E., & Talbot, C. 2019, PASA, 36, e010
Tiwari, V. 2022, ApJ, 928, 155
Tiwari, V., & Fairhurst, S. 2021, ApJL, 913, L19
van Son, L. A. C., de Mink, S. E., Callister, T., et al. 2022a, ApJ, 931, 17
van Son, L. A. C., de Mink, S. E., Renzo, M., et al. 2022b, ApJ, 940, 184
Veitch, J., Raymond, V., Farr, B., et al. 2015, PhRvD, 91, 042003
Wong, K. W. K., & Cranmer, M. 2022, arXiv:2207.12409
Woosley, S. E. 2017, ApJ, 836, 244
Woosley, S. E. 2019, ApJ, 878, 49
Woosley, S. E., & Heger, A. 2015, in Very Massive Stars in the Local Universe (Astrophysics and Space Science Library) Vol 412, ed. J. S. Vink (Cham: Springer)
Wysocki, D., & O'Shaughnessy, R. 2021, PopModels O3a APS April 2021, GitLab, https://gitlab.com/dwysocki/pop-models-o3a-aps-april-2021
Ye, C., & Fishbach, M. 2022, ApJ, 937, 73
Zevin, M., Bavera, S. S., Berry, C. P. L., et al. 2021, ApJ, 910, 152
Zevin, M., & Holz, D. E. 2022, ApJL, 935, L20
Zevin, M., Pankow, C., Rodriguez, C. L., et al. 2017, ApJ, 846, 82
Zevin, M., Spera, M., Berry, C. P. L., & Kalogera, V. 2020, ApJL, 899, L1