The Astrophysical Journal, 926:34 (16pp), 2022 February 10 https://doi.org/10.3847/1538-4357/ac3978

Probing Extremal Gravitational-wave Events with Coarse-grained Likelihoods

Reed Essick[1], Amanda Farah[2], Shanika Galaudage[3,4], Colm Talbot[5,6,7], Maya Fishbach[8,10], Eric Thrane[3,4], and Daniel E. Holz[2,9]

[1] Perimeter Institute for Theoretical Physics, 31 Caroline Street North, Waterloo, ON N2L 2Y5, Canada; reed.essick@gmail.com
[2] Department of Physics, University of Chicago, Chicago, IL 60637, USA
[3] School of Physics and Astronomy, Monash University, Clayton, VIC 3800, Australia
[4] OzGrav: The ARC Centre of Excellence for Gravitational Wave Discovery, Clayton, VIC 3800, Australia
[5] LIGO Laboratory, California Institute of Technology, Pasadena, CA 91125, USA
[6] LIGO Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[7] Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[8] Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA) and Department of Physics and Astronomy, Northwestern University, 1800 Sherman Ave., Evanston, IL 60201, USA
[9] Department of Astronomy and Astrophysics, Enrico Fermi Institute, and Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA
Received 2021 September 1; revised 2021 November 5; accepted 2021 November 11; published 2022 February 9

Abstract

As catalogs of gravitational-wave transients grow, new records are set for the most extreme systems observed to date. The most massive observed black holes probe the physics of pair-instability supernovae while providing clues about the environments in which binary black hole systems are assembled. The least massive black holes, meanwhile, allow us to investigate the purported neutron star–black hole mass gap, and binaries with unusually asymmetric mass ratios or large spins inform our understanding of binary and stellar evolution. Existing outlier tests generally implement leave-one-out analyses, but these do not account for the fact that the event being left out was by definition an extreme member of the population. This results in a bias in the evaluation of outliers. We correct for this bias by introducing a coarse-graining framework to investigate whether these extremal events are true outliers or whether they are consistent with the rest of the observed population. Our method enables us to study extremal events while testing for population model misspecification. We show that this ameliorates biases present in the leave-one-out analyses commonly used within the gravitational-wave community. Applying our method to results from the second LIGO–Virgo transient catalog, we find qualitative agreement with the conclusions of Abbott et al. GW190814 is an outlier because of its small secondary mass. We find that neither GW190412 nor GW190521 is an outlier.

Unified Astronomy Thesaurus concepts: Gravitational waves (678); Bayesian statistics (1900); Hierarchical models (1925); Catalogs (205); Gravitational wave astronomy (675)

1. Introduction

As catalogs of gravitational-wave (GW) sources observed with the Advanced LIGO (Aasi et al. 2015) and Virgo (Acernese et al. 2014) interferometers continue to grow, our knowledge of the population of compact objects is continually refined. The most recent update from the LIGO–Virgo–KAGRA (LVK) collaborations (GWTC-2; Abbott et al. 2021b) brings to light several interesting features within the distributions of masses and spins of compact objects in coalescing binary systems (Abbott et al. 2021a). In particular, GWTC-2 set new records for the largest black hole mass (GW190521; Abbott et al. 2020d), smallest black hole mass (GW190814; Abbott et al. 2020a),[11] and most asymmetric mass ratios (GW190814 and GW190412; Abbott et al. 2020a, 2020c). Of immediate interest is whether these objects are merely the most extreme events observed from a single population: the most extreme examples in a catalog become more extreme as the size of the catalog grows (Fishbach et al. 2020). Alternatively, these events may be inconsistent with the population inferred from the rest of the detected events and thus are *true outliers*. The observation of such outliers could suggest the first example from an as-of-yet-unmodeled or entirely new (sub)population. However, it may simply indicate that the current phenomenological population models are a poor description of nature.

The interpretation of the most extreme GW events has significant astrophysical implications. The most massive binary black hole (BBH) events probe the pair-instability supernova (PISN) mass gap (Fishbach & Holz 2017; Talbot & Thrane 2018), a theoretically proposed dearth of black holes between ∼50 and 120 $M_\odot$ (Heger & Woosley 2002). Observing BBH systems near the pair-instability gap can inform our knowledge of nuclear reaction rates (Farmer et al. 2020), beyond standard model physics (Croon et al. 2020; Baxter et al. 2021), or the boundaries of mass gaps in general (Fishbach & Holz 2020b; Edelman et al. 2021; Ezquiaga & Holz 2021; Nitz & Capano 2021) and applications thereof (e.g., Farr et al. 2019). At the other extreme, BBHs with small component masses can inform our knowledge of supernova physics (e.g., Fryer & Kalogera 2001; Belczynski et al. 2012; Zevin et al. 2020).

Several authors have posited that the most massive black holes of GWTC-2, which appear to sit in the PISN mass gap, form a separate population from the "main population" observed to date, invoking formation scenarios such as hierarchical mergers (e.g.,

[10] NASA Hubble Fellowship Program Einstein Postdoctoral Fellow.
[11] It is possible that the secondary object in GW190814 is actually an unusually massive neutron star (Essick & Landry 2020).

Abbott et al. 2020b; Kimball et al. 2021; Gerosa & Fishbach 2021; Tagawa et al. 2021) or primordial black holes (e.g., De Luca et al. 2021; Franciolini et al. 2021). Indeed, in order to explain the wide range of observed BBH properties, many authors have argued that multiple formation channels are active and that the BBH population consists of multiple subpopulations (Ng et al. 2021; Abbott et al. 2021a; Zevin et al. 2021). More than anything else, the variety of interpretations of the GWTC-2 events highlight the excitement within the GW community as new discoveries are routinely made whenever more data are recorded. However, they also emphasize the need for thoughtful consideration of the methods employed to assess whether individual events are consistent with existing population models.

Motivated by the "leave-one-out" consistency checks in Abbott et al. (2019a, 2021a), Fishbach et al. (2020), and elsewhere, we introduce a general procedure to investigate the effect of individual events on inferred populations. Specifically, we derive a "coarse-grained" analysis that retains, in a controlled way, only a subset of the total information available about some detected events. We are then able to determine whether the population inferred from these "coarse-grained" data is consistent with the population inferred from the original data.

In particular, previous approaches compared populations inferred with all N events in a catalog to populations inferred with only $N-1$ events, arguing that if these were similar then the event that was "left out" is consistent with the rest of the population, as its inclusion does not significantly change the inferred population. If the contrary were true, the event would be considered a true outlier. However, such tests do not account for the way in which the excluded event was selected. It was left out specifically because it was extremal in some dimension, and this selection may artificially inflate the apparent significance of differences in inferred populations, particularly in the presence of sharp boundaries within the population model. Our approach improves on this by self-consistently accounting for how the selected events are chosen in a controlled way while clearly defining the subset of the available information that is retained. We show that our approach naturally reproduces previous techniques when a random event is excluded from the analysis. It is precisely because the excluded events are typically not chosen at random that previous techniques can introduce biases.

There is already a healthy literature proposing leave-one-out analyses as tests for model misspecification within Bayesian inference. For example, Vehtari et al. (2017) construct a cross-validation likelihood from analyses that leave out each event in a catalog one at a time. This approach attempts to limit the possible biases associated with leaving out only extremal events, similar to the motivation for our coarse-grained approach. Other authors have considered tempering the likelihood in order to combat model misspecification, which is at times referred to as coarsening (Miller & Dunson 2019). In this approach, the likelihood is raised to the power of an inverse temperature $\beta = 1/T \in (0,\ 1]$, thereby artificially inflating statistical uncertainties. Several procedures exist to select β (e.g., Miller & Dunson 2019; Thomas & Corander 2019, and references therein), all of which attempt to optimize the balance between systematic error from model misspecification and additional statistical error from tempered likelihoods. Our coarse-grained likelihoods are similar in spirit but differ in that we eschew the use of ad hoc tempering in favor of marginalizing over the data and parameters from an individual event subject to a precisely specified (but looser) constraint on the event's parameters. That is, we specify the size and placement of the "coarse grain" used to approximate the event's likelihood.

Using this coarse-grained likelihood, we revisit several astrophysically interesting events from GWTC-2 that were discussed in detail in Abbott et al. (2021a). To wit, we demonstrate the following:

1. GW190814, with a secondary mass of $m_2 \sim 2.6\,M_\odot$ and mass ratio $q = m_2/m_1 \sim 0.1$, is an outlier in m_2 (too small to be consistent with the main BBH population) but is not an outlier in mass ratio.
2. GW190412 is not an outlier, and its mass ratio ($q \sim 0.28$; Abbott et al. 2020c) is in the tail of the main BBH population.
3. GW190521 is not an outlier with respect to the main BBH population under the preferred phenomenological mass models considered in Abbott et al. (2021a). It is only marginally inconsistent under the simplest mass model considered, which is disfavored for other reasons as well.

While we believe that the quantitative details of our analysis improve on previous methods, all of our resulting astrophysical conclusions are in agreement with Abbott et al. (2021a).

The remainder of this paper is organized as follows. We derive our coarse-grained leave-one-out formalism from first principles in Section 2 and explore a toy model in Section 3 to gain intuition. Readers only interested in our astrophysical conclusions can skip directly to Section 4, where we apply the method to public LVK data. We conclude in Section 5.

2. Coarse-grained Leave-one-out Likelihoods

We derive a procedure for coarse-graining our inference of the population in Section 2.1, while in Section 2.2 we discuss the implications of different possible procedures to choose the amount of information retained in the coarse-grained inference. We describe how to quantify these consistency tests with a single statistic in Section 2.3.

2.1. Derivation of Coarse-grained Inference

To begin, let us assume we have N events that are each described by a set of m parameters $\theta \in \mathcal{R}^m$. We further assume that these events belong to the same astrophysical population described by the hyperparameters Λ_0:

$$\theta_i \sim p(\theta|\Lambda_0) \quad \forall \quad i. \tag{1}$$

If we write our model for the differential Poisson number density of signals in the universe as $d\mathcal{N}/d\theta = Rp(\theta|\Lambda)$, our joint distribution over the data $\{D_i\}$, single-event parameters $\{\theta_i\}$, hyperparameters, and the rate R is then

$$p(\{D_i\};\ \{\theta_i\};\ \Lambda,\ R) = p(\Lambda)p(R)R^N e^{-R\mathcal{E}(\Lambda)} \prod_{i=1}^{N} p(D_i|\theta_i)p(\theta_i|\Lambda)\Theta(\rho(D_i) \geqslant \rho_{\rm thr}), \tag{2}$$

where Θ is the Heaviside function, $\rho(D)$ is the detection statistic used to select events for the catalog with threshold $\rho_{\rm thr}$,

and

$$\begin{aligned}\mathcal{E}(\Lambda) &= \int d\theta\, p(\theta|\Lambda) \int dD\, p(D|\theta)\Theta(\rho(D) \geqslant \rho_{\rm thr}) \\ &= \int d\theta\, p(\theta|\Lambda) P(\mathrm{det}|\theta) \\ &= P(\mathrm{det}|\Lambda) \end{aligned} \tag{3}$$

is the expected fraction of events that are detectable within a population. Note that Equation (2) is neither a likelihood function nor a posterior distribution, but instead is a joint distribution for both the data and the parameters. Furthermore, if we assume a uniform-in-log prior for the rate ($p(R) \sim 1/R$) and marginalize, we obtain

$$p(\{D_i\};\, \{\theta_i\};\, \Lambda) = p(\Lambda) \prod_{i=1}^{N} \left[\frac{p(D_i|\theta_i) p(\theta_i|\Lambda) \Theta(\rho(D_i) \geqslant \rho_{\rm thr})}{\mathcal{E}(\Lambda)} \right]. \tag{4}$$

We note that the Θ in each factor within the product does not impact the inference, as it is guaranteed to be 1; the data are axiomatically detectable for detected events. However, these factors are important in what follows.

If we wish to construct a posterior for only Λ, we marginalize over $\{\theta_i\}$ and condition on the observed data to obtain

$$p(\Lambda|\{D_i\}) \propto p(\Lambda) \prod_{i=1}^{N} \left[\frac{\int d\theta_i\, p(D_i|\theta_i) p(\theta_i|\Lambda)}{\mathcal{E}(\Lambda)} \right]. \tag{5}$$

All these expressions have become commonplace within the GW community, and we refer the reader to the many reviews in the literature for more details (e.g., Mandel et al. 2016; Thrane & Talbot 2019; Vitale et al. 2020).

We are particularly interested in the effect that subsets of events have on the inferred population and whether those effects can be used to determine if particular events are outliers. However, if we simply remove potential outliers, we may bias the inferred population by preferentially removing the most extreme events. This, in turn, may artificially inflate the significance of any changes in the inferred population. As such, we introduce a coarse-graining procedure to replace traditional "leave-one-out" analyses. This procedure retains only a limited amount of information about a particular event. In this way, the analysis can naturally account for how the event of interest was selected (e.g., as the most extreme event in some dimension) while still discarding as much information about that event as possible.

To wit, let us assume that the jth event is almost certainly within some region of parameter space $\mathcal{S}$ so that

$$\int d\theta_j p(\theta_j|D_j,\, \Lambda) \Theta(\theta_j \in \mathcal{S}) \geqslant 1 - \epsilon \quad \forall \quad \{\Lambda:\, p(\Lambda) \neq 0\}. \tag{6}$$

In other words, for any population described by Λ with nonvanishing support in the hyperprior $p(\Lambda)$, the inferred posterior on the event parameters θ_j is contained within the region $\mathcal{S}$ with high $(1-\epsilon)$ credibility. We are typically interested in the case $\epsilon \ll 1$, which is to say we are confident that event j is contained within $\mathcal{S}$. There is a uniquely defined smallest $\mathcal{S}$ that satisfies this requirement, which in general depends on the choice of ϵ (smaller ϵ require larger $\mathcal{S}$). Additionally, we only assert that $\mathcal{S}$ is at least as large as the minimal region; it may be larger. Choosing a larger region corresponds to retaining less information about the jth event.

At times, it is possible to identify individual events that are clearly separated from the rest of the population, in which case it may be straightforward to identify an appropriate choice for $\mathcal{S}$. An example could be defining $\mathcal{S}$ as the region with masses smaller than the minimum mass observed in a set. However, more complicated boundaries are possible, as hypersurfaces in multidimensional spaces could divide events along a nontrivial slice that depends on multiple parameters simultaneously. For example, we may define $\mathcal{S}$ to be a region with a secondary mass or mass ratio smaller than the corresponding minima observed in a set (see Section 4.1). We discuss several procedures to choose $\mathcal{S}$ in Section 2.2.

Of course, an extremal event is not necessarily inconsistent with the population determined by the other $N-1$ events; some event is always the most extreme of any set. Indeed, the jth event may still be drawn from the same $p(\theta|\Lambda_0)$ and simply be the most extreme example in the catalog. We therefore test the null hypothesis that all N events are drawn from the same distribution by comparing the inferred population when we include all N events in the population inferred from $N-1$ events while accounting for the knowledge that $\theta_j \in \mathcal{S}$. If the population inferred from the coarse-grained analysis is inconsistent with the full data set, we reject the null hypothesis that the jth event is drawn from the same distribution as the other $N-1$ events. With this method, we can be assured that we do not overestimate the significance of differences in the inferred populations because we include knowledge that the jth event was selected in a particular way.

We now consider how to self-consistently limit the information about the jth event within our inference. Because $\theta_j \in \mathcal{S}$ almost surely ($\epsilon \ll 1$), we can add a term to Equation (4) without affecting the overall inference for Λ:

$$\begin{aligned} p(\{D_i\};\, \{\theta_i\};\, \Lambda) &\simeq p(\Lambda) \\ &\times \left(\frac{p(D_j|\theta_j) p(\theta_j|\Lambda) \Theta(\rho(D_j) \geqslant \rho_{\rm thr}) \Theta(\theta_j \in \mathcal{S})}{\mathcal{E}(\Lambda)} \right) \\ &\times \prod_{i \neq j}^{N-1} \left[\frac{p(D_i|\theta_i) p(\theta_i|\Lambda) \Theta(\rho(D_i) \geqslant \rho_{\rm thr})}{\mathcal{E}(\Lambda)} \right], \end{aligned} \tag{7}$$

where we have included an additional factor of $\Theta(\theta_j \in \mathcal{S})$ since this is 1 almost everywhere there is posterior support for θ_j (to within the precision specified in Equation (6)). This is completely analogous to the way in which terms representing the detectability of data, i.e., $\Theta(\rho(D) \geqslant \rho_{\rm thr})$, are always 1 for the detected events. We have also replaced "=" with "$\simeq$" to denote that strict equality only holds in the limit $\epsilon \to 0$. However, in what follows we assume that ϵ is small enough to be negligible, and we retain the "=" for simplicity.

Furthermore, the only information we wish to retain about the jth event is that it almost certainly is within $\mathcal{S}$, and therefore we marginalize over both D_j and θ_j to effectively forget

everything else about the event. This yields

$$\begin{aligned}\int dD_j d\theta_j\, p(D_j|\theta_j)p(\theta_j|\Lambda) &\\ \times \Theta(\rho(D_j) \geqslant \rho_{\rm thr})\Theta(\theta_j \in \mathcal{S}) &\\ = \int d\theta_j\, \Theta(\theta_j \in \mathcal{S})p(\theta_j|\Lambda) &\\ \times \int dD_j p(D_j|\theta_j)\Theta(\rho(D_j) \geqslant \rho_{\rm thr}) &\\ = \int d\theta_j\, \Theta(\theta_j \in \mathcal{S})p(\theta_j|\Lambda)P(\det|\theta_j) &\\ = P(\det,\theta_j \in \mathcal{S}|\Lambda), \end{aligned} \tag{8}$$

which is just the probability that the event was detected and had parameters within $\mathcal{S}$ given the underlying population model. We additionally note that this term only appears in Equation (7) within a ratio, and the divisor of that ratio is $\mathcal{E}(\Lambda) = P(\det|\Lambda)$. Therefore, the coarse-graining procedure yields an overall factor of

$$\frac{P(\det,\theta_j \in \mathcal{S}|\Lambda)}{P(\det|\Lambda)} = P(\theta_j \in \mathcal{S}|\det,\Lambda). \tag{9}$$

This has the appealing interpretation that the only information retained about the jth event is the probability that it belongs to a particular part of parameter space (our "coarse grain") given that it was detected and came from a particular population.[12] We discuss implications and implicit assumptions made by the choice of $\mathcal{S}$ in more detail in Section 2.2. Briefly, we note that most "agnostic" procedures for defining $\mathcal{S}$ will not depend on D_j or θ_j, and so we explicitly write $\mathcal{S} = \mathcal{S}(\{\theta_{i\neq j}\})$ below.

Putting everything together, we marginalize the coarse-grained likelihood over $\{\theta_{i\neq j}\}$ and condition on $\{D_{i\neq j}\}$ to obtain a coarse-grained posterior for Λ,

$$\begin{aligned} p(\Lambda|\{D_{i\neq j}\},\, \rho(D_j) \geqslant \rho_{\rm thr};\, \theta_j \in \mathcal{S}(\{\theta_{i\neq j}\})) &\\ \propto p(\Lambda) \int \left(\left[\prod_{i\neq j}^{N-1} d\theta_i \right] \left[\prod_{k\neq j}^{N-1} \frac{p(D_k|\theta_k)p(\theta_k|\Lambda)}{\mathcal{E}(\Lambda)} \right] \right. &\\ \left. \times\, P(\theta_j \in \mathcal{S}(\{\theta_{i\neq j}\})|\det,\Lambda) \right). \end{aligned} \tag{10}$$

We note that the marginalization over $\{\theta_{i\neq j}\}$ does not factor as nicely as it does in Equation (5) because $\mathcal{S}(\{\theta_{i\neq j}\})$ may depend on all the events (except the jth event) and is included within the integral. Methods to calculate Equation (10) efficiently when one already has access to samples from $p(\Lambda|\{D_{i\neq j}\})$ or $p(\Lambda|\{D_i\})$ are discussed in Appendix A.

2.2. Implications from the Choice of $\mathcal{S}$

Section 2.1 presents the coarse-grained inference for Λ based on knowledge that $\theta_j \in \mathcal{S}$. This holds for arbitrary $\mathcal{S}$ as long as Equation (6) is satisfied; the choice of $\mathcal{S}$ is up to the analyst. We now consider the implications of different choices for $\mathcal{S}$ and what assumptions they represent.

To begin, if θ_j is well determined by D_j and the minimum allowable $\mathcal{S}$ is small (particularly in comparison to the extent of the prior $p(\theta|\Lambda)$), the coarse-graining procedure can still retain most, if not all, of the relevant information about the jth event. In this way, defining $\mathcal{S}$ in relation to the minimum allowable $\mathcal{S}$ implicitly assumes some knowledge of D_j, which is undesirable for a leave-one-out analysis. To avoid this, we instead recommend choosing $\mathcal{S}$ based on only a priori theoretical predictions of astrophysical interest or the data from the $N-1$ other events. In general, this implies that one should write $\mathcal{S} = \mathcal{S}(\{\theta_{i\neq j}\})$ to explicitly show that it does not depend on either the data or parameters from the jth event, as we do within Equation (10).

One possible data-driven procedure for choosing $\mathcal{S}$ is particularly appealing in the special case where the parameters of the jth event are cleanly separated from the rest of the events. If the jth event is the most extreme event in some dimension x, we can simply define $\mathcal{S}(\{x_{i\neq j}\})$: $x \geqslant \max_{i\neq j}\{x_i\}$ as the region where x is larger than the second-largest detected event.[13] This choice is natural in the sense that it only depends on $\{D_{i\neq j}\}$ and is as generous as possible subject to the knowledge that the jth event is extremal; it does not retain any information about how much larger x_j is than $\max_{i\neq j}\{x_i\}$.

Such a choice allows an analyst to pose questions such as, "how big should the largest individual event in a catalog of N events be given the observation of $N-1$ events and the knowledge that one was larger?" Comparing the predicted largest event to what was actually observed naturally defines a metric that can be used as a quantitative consistency check (see Section 2.3). Indeed, the choice of $\mathcal{S}$ precisely specifies the information used when computing p-values for the null hypothesis that all N events in a catalog were drawn from the same population. Nonetheless, it is important to remember that this procedure is not unique, just as the definition of a null hypothesis is not unique, and other choices of $\mathcal{S}$ may be able to more naturally answer other questions.

In this vein, one might also consider choosing $\mathcal{S}$ based on prior theoretical motivations, so long as the selected region is compatible with Equation (6). This can provide a perfectly natural way to define alternative tests of the null hypothesis and is particularly relevant when the events are not cleanly separated. For example, even before observing any data, we may identify the region $m_1 > 50\,M_\odot$ as interesting from the standpoint of PISN theory and scrutinize any events that fall in this region as potential outliers.

We note that the inability to define $\mathcal{S}$ based on the next most extreme event only arises when the jth event is not cleanly separated from the other events, and therefore it is not unambiguously the most extreme event. One may take that ambiguity itself as evidence that the event cannot be an outlier and therefore argue that our machinery may not be needed in such situations. Nonetheless, we can still construct a perfectly self-consistent coarse-grained inference even when the minimal $\mathcal{S}$ surrounding the jth event's parameters overlaps significantly with the inferred parameters of the other $N-1$ events. The coarse-grained event need not actually be extremal.

As a limiting case, we note that choosing $\mathcal{S}$ that spans the entire parameter space ($\mathcal{S} \to \mathcal{R}^m$) implies that $P(\theta_j \in S|\det,\Lambda) \to 1\ \forall\ \Lambda$. That is, if we know nothing about the parameters of the jth event, then the coarse-grained inference is equivalent to neglecting that event altogether and performing the "standard" inference for Λ with the other $N-1$ events.[14] This is intuitive in

[12] Note that Equation (9) refers to the probability that a detected event from a population came from $\mathcal{S}$, whereas Equation (6) refers to the probability that a specific event (the jth observed event, which produced data D_j) came from $\mathcal{S}$.

[13] One could equivalently consider the smallest event.

[14] The corresponding inference of the rate would still correctly account for the fact that we detected N events in total, which is not necessarily true of other leave-one-out analyses.

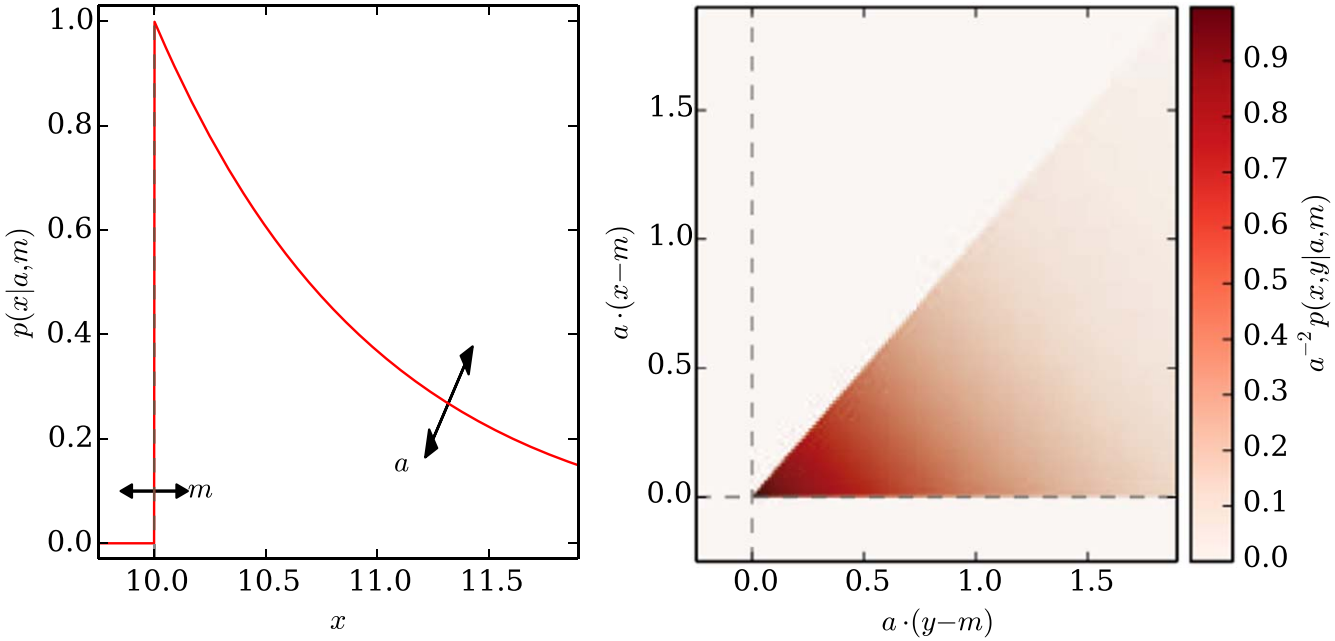

Figure 1. Toy population model with a sharp lower limit. We show the distribution of single parameters (Equation (12)), annotated to show how the hyperparameters affect the shape of the distribution (left), and the joint distribution (Equation (13)) from which individual events are drawn to form our synthetic catalogs (right).

that, if we picked an event to coarse-grain at random without first considering its parameters, then the smallest allowable $\mathcal{S}$ that was certain to contain the event would be the entire (detectable) parameter space. At the same time, excluding an event at random should not introduce any biases within the inference for Λ, and one expects to perform the standard inference with only $N-1$ events. We see, then, that performing a standard inference with $N-1$ events after excluding an extremal event, *without* incorporating the knowledge that the excluded event was extremal, is not a self-consistent procedure because it does not correctly reflect the data selection procedure. This inconsistency leads to biases, which we demonstrate with a toy model in Section 3.

2.3. Quantifying Inconsistencies with Inferred Populations

In order to test the null hypothesis that the event in question is consistent with the $N-1$ other events, we follow a similar procedure to Fishbach et al. (2020) and construct a p-value statistic. We marginalize over both the uncertainty in the population hyperparameters, inferred from the coarse-grained inference, and the measurement uncertainty in the parameters of the N observations. For each hyperposterior sample Λ from the coarse-grained inference, we do the following:

1. Draw N synthetic detected events from the corresponding population described by those hyperparameters.
2. Reweigh the N observed events' posteriors for $\{\theta_i\}$ to what would be obtained under the population prior described by that Λ. For the $N-1$ events that were not excluded, this recovers their marginal posterior for $\{\theta_{i\neq j}\}$ under the joint distribution of Equation (7), while for the excluded event we obtain the population-informed posterior for θ_j.
3. Draw one $\theta_{\rm obs}$ sample for each observed event.
4. Compare the most extreme $\theta_{\rm syn}$ out of the N synthetic events to the most extreme $\theta_{\rm obs}$.

Repeating this procedure for many hyperposterior samples Λ, the p-value is obtained as the fraction of synthetic catalogs that produce an event that is at least as extreme as the most extreme observed event. Although such statistics are not necessarily uniformly distributed even under the null hypothesis (see, e.g., Seth et al. 2019), if the resulting marginalized p-value is small, we reject the null hypothesis and conclude that the excluded event is inconsistent with the main population.

Fishbach et al. (2020) estimated p-values by scattering the maximum likelihood estimates by synthetic noise realizations and then comparing the maximum likelihood from the observed event to the synthetic distribution under the null hypothesis. Our approach differs in that we do not scatter the maximum likelihood estimate of our synthetic detections, but instead use the true values of the detected events. However, we still marginalize over the actual uncertainty in the observed events' true parameters stemming from detector noise. Both procedures, then, account for measurement uncertainty but answer slightly different questions.

3. Toy Model

We first investigate a simple toy model to gain an intuition for how our coarse-graining formalism impacts population inferences. Specifically, we investigate the impact of defining $\mathcal{S}$ based on multiple single-event parameters in the presence of sharp edges within a population model (e.g., mass gaps or spin cutoffs). Of particular interest is the impact on our uncertainty in the inferred location of the sharp edges (see Farr et al. 2019; Ezquiaga & Holz 2021).

Figure 1 shows our population model. We assume that individual events are described by two real parameters

$$\begin{aligned} x_i, y_i &\sim p(x, y|a, m) \\ &= p(x|a, m)p(y|a, m)\Theta(m \leqslant x \leqslant y), \end{aligned} \tag{11}$$

where

$$p(x|a, m) = ae^{-a(x-m)}\Theta(m \leqslant x). \tag{12}$$

In this model, objects are independently drawn from the same distribution, which has a sharp cutoff at $x = m$, and are then

randomly paired subject to an arbitrary labeling scheme ($x \leqslant y$). This implies

$$p(x, y|a, m) = 2a^2 e^{-a(x+y-2m)}\Theta(m \leqslant x \leqslant y). \tag{13}$$

In this model, m controls the smallest allowed value within the population and a controls the spread in values; larger a imply faster exponential decay and more tightly clustered values. Furthermore, we assume that all events are observable ($\mathcal{E}(a, m) = 1 \ \forall a, m$) and have vanishingly small observational uncertainties ($p(x, y|D_i) \sim \delta(x - x_i)\delta(y - y_i)$) for simplicity. This also implies that there is no ambiguity in which event is the most extreme in any dimension. The full hyperposterior is then

$$\begin{aligned} p(a, m|\{x_i, y_i\}) &\propto p(a, m) \prod_i^N p(x_i, y_i|a, m) \\ &\propto p(a, m)(2a^2)^N e^{-a\sum_i^N (x_i+y_i-2m)} \\ &\times \Theta(m \leqslant \min_i\{x_i\}) \end{aligned} \tag{14}$$

with hyperprior $p(a, m)$. For concreteness, we assume

$$p(a) = \frac{1}{A\Gamma(\alpha + 1)}\left(\frac{a}{A}\right)^\alpha e^{-a/A} \tag{15}$$

$$p(m) = \frac{1}{M\Gamma(\mu + 1)}\left(\frac{m}{M}\right)^\mu e^{-m/M}, \tag{16}$$

which render the hyperposterior analytically tractable. Figures 2 and 3 consider the limits $\alpha, \mu \to 0$ and $A, M \to \infty$ to obtain uninformative hyperpriors.

We also consider the coarse-grained hyperposterior, which takes the form

$$\begin{aligned} p(a, m|\{x_{i\neq j}, y_{i\neq j}\}, (x_j, y_j) \in \mathcal{S}) \\ \propto p(a, m|\{x_i, y_i\})\frac{P(x, y \in \mathcal{S}|a, m)}{p(x_j, y_j|a, m)}, \end{aligned} \tag{17}$$

where

$$\begin{aligned} P(x, y \in \mathcal{S}|a, m) \\ \equiv \int_{\mathcal{S}} dxdy\ 2a^2 e^{-a(x+y-2m)}\Theta(m \leqslant x \leqslant y). \end{aligned} \tag{18}$$

Because there is no ambiguity in which event is the most extreme, we choose $\mathcal{S}$ based on the parameters of the other $N - 1$ events. We consider two special cases: excluding the event with the smallest x (Section 3.1) and excluding the event with the smallest $q \equiv x/y$ (Section 3.2). In what follows, we consider catalogs of 10 events in total, characteristic of GWTC-1 (Abbott et al. 2019b). Appendix B explores how our toy models scale with larger catalogs.

3.1. Excluding the Smallest x

If we exclude the event with the smallest x, we obtain $\mathcal{S}$: $x \leqslant \min_{i\neq j}\{x_i\} \equiv \mathcal{M}$ and

$$\begin{aligned} P(x, y \in \mathcal{S}|a, m) &= P(x \leqslant \mathcal{M}|a, m) \\ &= \int_m^{\mathcal{M}} dx \int_x^\infty dy\ 2a^2 e^{-a(x+y-2m)} \\ &= (1 - e^{-2a(\mathcal{M}-m)})\Theta(m \leqslant \mathcal{M}). \end{aligned} \tag{19}$$

From this, we can immediately write down the full hyperposterior with Equation (17).

In our simple model, excluding the event with the smallest x primarily impacts our knowledge of m, the minimum allowed value within the population. Figure 2 shows two examples with synthetic data, one in which all simulated events are drawn from the same population (our null hypothesis) and one in which a single event is drawn from a different population centered at $x \ll m$ (a true outlier) while the other $N - 1$ events are drawn from the original population. For each example, we show the inferred hyperposterior using all N events (Equation (14)), the coarse-grained hyperposterior (Equations (17) and (19)), and the inferred posterior from Equation (14) when we use only $N - 1$ events and do *not* account for the fact that we excluded the event with the smallest x.

If the null hypothesis is true, we see that the coarse-grained hyperposterior agrees well with the hyperposterior that uses all $N = 10$ events. The coarse-graining procedure correctly accounts for the additional probability associated with detecting an event with small x. Conversely, the hyperposterior using $N - 1$ events without the coarse-graining correction is biased toward higher m. Indeed, this bias could lead to the erroneous identification of the smallest event as inconsistent with the main population.

In addition to individual realizations of synthetic catalogs, Figure 2 also shows the cumulative distributions of the total probability from the region assigned a posterior probability $p(\Lambda|\{D\}) \geqslant p(\Lambda_0|\{D\})$, the posterior at the true hyperparameters. Correct coverage corresponds to diagonal lines, and the shaded gray regions demonstrate the size of expected 1σ, 2σ, and 3σ fluctuations from the finite number of trials performed. When the null hypothesis is true, we see that the coarse-grained inference is unbiased; it agrees well with the full N-event hyperposterior and has correct coverage. Conversely, the $(N - 1)$-event hyperposterior that does not include the coarse-graining correction does not have correct coverage; the true population parameters are systematically assigned posterior probabilities that are too low.

When the null hypothesis is false and the smallest event was not drawn from the same population as the other $N - 1$ events, we see markedly different behavior. Here, the full N-event hyperposterior is biased to significantly lower m while both of the $(N - 1)$-event inferences are much less affected. The $(N - 1)$-event inference that does not include the coarse-graining correction is unbiased and has correct coverage in this case. It correctly excludes the extremal event from the inference of the main population. The coarse-grained $(N - 1)$-event hyperposterior is biased when the null hypothesis is incorrect, as it incorrectly assumes that the jth event is drawn from the same population as the other $N - 1$ events, but it is much less biased than the full N-event result. Indeed, it appears to have nearly correct coverage.

This suggests the following rule of thumb: If the null hypothesis is correct, the full N-event hyperposterior should be very similar to the $(N - 1)$-event coarse-grained hyperposterior. However, if the null hypothesis is incorrect, the coarse-grained hyperposterior is likely to be more similar to the $(N - 1)$-event hyperposterior that does not contain the coarse-graining correction. However, this may be violated in practice (see Section 4.3), and we suggest that decisions be based on quantitative assessments like those proposed in Section 2.3. Furthermore, in both cases we note that one should not use the coarse-grained hyperposterior as the "final inferred population." Even though it consistently provides a reasonable estimate of the uncertainty in the population, it is only a useful diagnostic tool to determine which of the other hyperposteriors we believe. We also suggest

that estimates of p-values be performed with the coarse-grained hyperposterior (see Section 3.3).

3.2. Excluding the Smallest q ≡ x/y

We additionally investigate the impact of discarding the event with the smallest q, defining $\mathcal{S}$: $q \leqslant \min_{i\neq j}\{q_i\} \equiv \mathcal{Q}$. With the coordinate change $q = x/y$ and $r = x + y$, we obtain

$$\begin{aligned} P(x, y \in \mathcal{S}: q \leqslant \mathcal{Q}|a, m) &= 2a^2 e^{2ma} \int_0^{\mathcal{Q}} dq\ (1+q)^{-2} \int_{m(1+q)/q} dr\ r e^{-ar} \\ &= 2\left(\frac{\mathcal{Q}}{1+\mathcal{Q}}\right) e^{ma(\mathcal{Q}-1)/\mathcal{Q}} \end{aligned} \tag{20}$$

because $r = y(1+q) = x(1+q)/q \geqslant m(1+q)/q$.

Figure 3 summarizes our conclusions. In general, our inference of a is more affected than m when excluding the event with the smallest q. This is because a more closely controls the range of values supported in the population (larger a imply faster exponential decay and more concentrated samples). If the events are restricted to a narrow range, then they are more likely to have $q \sim 1$. For this reason, the $(N-1)$-event hyperposterior that does not include the coarse-graining correction is biased to larger a (larger q) when the null hypothesis is true. Generally, the coverage is better in all cases, which we attribute to the lack of sharp features for q in the population model. As in Section 3.1, we find that the coarse-grained hyperposterior is generally more robust against the presence (or absence) of true outliers than either of the other options.

3.3. Testing the Null Hypothesis with Coarse-grained p-values

Beyond the intuition developed in Sections 3.1 and 3.2, we wish to quantify the probability that the full set of N events are drawn from the same distribution (our null hypothesis). Because the coarse-grained hyperposterior provides a reasonable estimate of the true underlying population regardless of whether the null hypothesis is correct, we compute p-values that assume that individual extremal events are consistent with the population inferred within the coarse-grained inference following the procedure detailed in Section 2.3. Indeed, a primary motivation for our coarse-grained inference is to avoid accidentally biasing such p-values to higher significance by excluding extremal events, or to artificially lower their significance by computing p-values with the full N-event population analysis.

We investigate several astrophysical events from GWTC-2 in Section 4, but first we consider the toy models in Sections 3.1 and 3.2 in more detail. In particular, we are concerned with changes in the probability of making *type 1 errors* when the null hypothesis is true (incorrectly rejecting the null hypothesis, or a false positive). At the same time, we are also interested in changes in the probability of making *type 2 errors* when the null hypothesis is false (incorrectly failing to reject the null hypothesis, or a false negative).

Figure 4 shows the distribution of such p-values for different realizations of synthetic catalogs when we exclude events with the smallest x (Section 3.1). We immediately note that the $(N-1)$-event inference, which neglects the coarse-graining correction, always predicts smaller p-values than the coarse-grained inference in both cases. In this way, analysts could be tricked into claiming more significant tension than actually exists between the inferred model and the excluded event if they do not account for how the event was selected. Similarly, they may be more likely to accept the null hypothesis when it is false based on coarse-grained inferences. In either case, though, the p-values differ by only a factor of a few. This may be why we reach the same astrophysical conclusions as Abbott et al. (2021a) even though they did not include coarse-graining corrections; the size of the effect on p-values is nontrivial but still relatively modest.

4. Astrophysical Results

Using the coarse-grained inference described in Section 2 and our intuition from the toy models investigated in Section 3, we revisit several astrophysical events from GWTC-2. We are specifically interested in evidence that individual events are incompatible with the main BBH population (the phenomenological distribution that describes the majority of detected BBH systems) inferred in Abbott et al. (2021a). We consider GW190814 (Section 4.1), GW190412 (Section 4.2), and GW190521 (Section 4.3) in turn. Our analysis reweighs publicly available population hyperposterior samples (The LIGO Scientific Collaboration & The Virgo Collaboration 2020a); see Appendix A for more details.

In what follows, we focus on results with the POWERLAW +PEAK mass model from Talbot & Thrane (2018) and Abbott et al. (2021a). Unless otherwise noted (i.e., GW190521 in Section 4.3), astrophysical conclusions are unchanged when we assume different mass models. Furthermore, we also consider fixed $\mathcal{S}$ in each case. Specific choices for $\mathcal{S}$ are listed in each section and are motivated by either the other $N-1$ events or theoretical expectations for astrophysical systems.

4.1. GW190814 Is an Outlier in Secondary Mass

We begin by considering GW190814 (Abbott et al. 2020a; The LIGO Scientific Collaboration & The Virgo Collaboration 2020b), the BBH system with the smallest secondary mass and the most extreme mass ratio observed to date. Indeed, GW190814's secondary is so small ($m_2 \sim 2.6\,M_\odot$) that there has been significant discussion about whether it could have been a neutron star (see, e.g., Essick & Landry 2020), with the common consensus that the system is likely incompatible with a slowly rotating neutron star. For simplicity, we eschew the question of whether both components of GW190814 were actually black holes and instead focus on whether their masses are compatible with the distribution inferred from the rest of the BBH events in GWTC-2.

We adopt the mass models explored in Abbott et al. (2021a), all of which include a cutoff at low masses (albeit with variable degrees of sharpness) in much the same way as our toy model (Section 3). In some sense, then, the results in Figure 5 are directly comparable to Figure 2.

We define $\mathcal{S}$ for our analysis of GW190814 as follows:

$$\mathcal{S}_{\rm GW190814}: (m_2 \leqslant 5.0\,M_\odot)\ \text{OR}\ (q \leqslant 0.28). \tag{21}$$

The $m_2 \leqslant 5.0\,M_\odot$ boundary is chosen to match the median posterior estimate of GW190924, the event with the second-smallest secondary mass after GW190814. The $q \leqslant 0.28$ boundary is chosen to match the median posterior estimate of GW190412, the event with the second-smallest mass ratio after

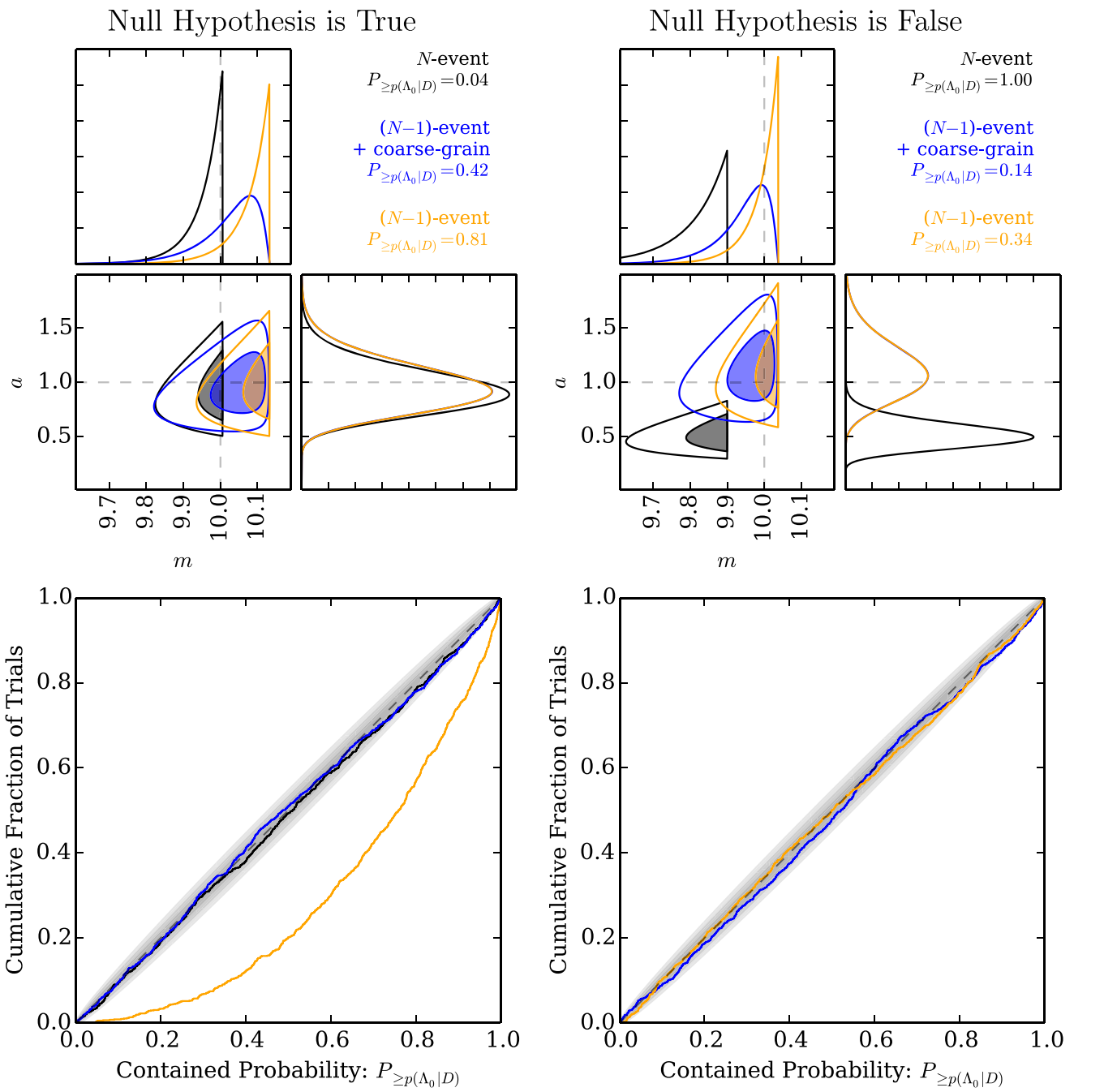

Figure 2. Toy model in which we exclude the event with the smallest x from catalogs of 10 events when all events are drawn from the same distribution (left) and $N-1$ events are drawn from the same distribution and there is an additional true outlier at $(x, y) = (9.9, 30)$ (right). Top row: example realizations of synthetic catalogs. Dashed lines correspond to the true hyperparameters. We see that the $(N-1)$-event coarse-grained inference ameliorates systematic biases in both the N-event and $(N-1)$-event inferences when their underlying assumptions are incorrect. Note that the marginal $(N-1)$-event hyperposteriors for a are almost identical. Bottom row: cumulative histograms of the posterior probability integrated over the region with $p(\Lambda|\{D\}) \geqslant p(\Lambda_0|\{D\})$. Diagonal lines indicate proper coverage, and shaded regions approximate expected 1σ, 2σ, and 3σ deviations from the counting uncertainty with the finite number of trials performed. We note that the $(N-1)$-event inference introduces large systematic biases (true hyperparameters are preferentially found in the tails of the hyperposterior) when the null hypothesis is true, while the N-event inference may not ever contain the true hyperparameters when the null hypothesis is false (the N-event curve is absent from the bottom right panel because the contained probability is always 1). However, in both cases, the $(N-1)$-event inference with coarse graining provides reliable posteriors, although we do expect them to be biased to some extent when the null hypothesis is false.

GW190814. Both GW190924 and GW190814 are highlighted in blue in Figure 5. This defines an "L-shaped" region in the (m_2, q) plane spanning the lowest values for both dimensions (see Figure 5). We again note that this choice of $\mathcal{S}$ is not unique, and one could instead choose to define $\mathcal{S}$ with bounds on only m_2 or only q. Defining $\mathcal{S}_{\rm GW190814}$ in terms of only m_2 or only q does not affect our conclusions, and defining $\mathcal{S}_{\rm GW190814}$ in this way allows us to be as agnostic as possible about GW190814 in some sense. Importantly, we note that, with this definition of $\mathcal{S}_{\rm GW190814}$, population models that do not have support for masses below $m_2 \leqslant 5.0\,M_\odot$ are still allowed as long as they support $q \leqslant 0.28$, and vice versa.

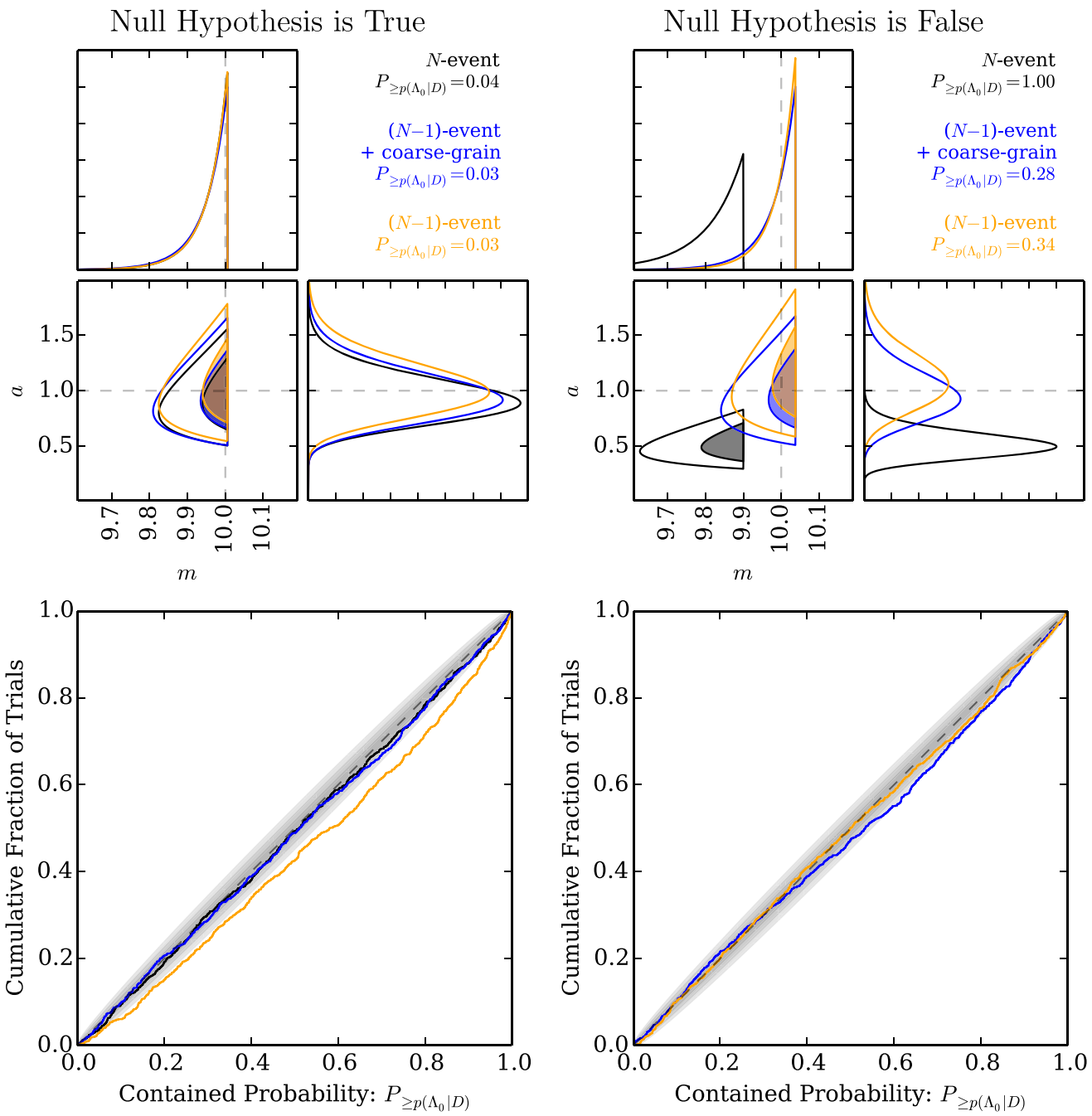

Figure 3. Analogous to Figure 2, except simulations exclude the event with the smallest $q \equiv x/y$. We note that biases from incorrect assumptions appear to be smaller when we exclude events with the smallest q compared to when we exclude events with the smallest x, which we attribute to the lack of a sharp feature in the population distribution over q. Nonetheless, we consistently observe incorrect coverage for the $(N-1)$-event hyperposterior when the null hypothesis is true at $\gtrsim 3\sigma$.

Figure 5 summarizes our conclusions. Specifically, we see that the $(N-1)$-event coarse-grained hyperposterior resembles the full N-event hyperposterior for β_q, the hyperparameter that controls the extent of the population model for q, while it more closely resembles the $(N-1)$-event hyperposterior that neglects the coarse-grained correction for both $m_{\rm min}$ and δ_m, which control the minimum mass allowed in the population. We note that $m_{\rm min}$ sets an absolute lower bound for the allowed masses within a population, and therefore we would expect a reasonable amount of probability that $m_{\rm min} \leqslant 3\,M_\odot$ in the $(N-1)$-event analyses if GW190814 was consistent with the population inferred from the other events. While $m_{\rm min} \leqslant 3\,M_\odot$ is not excluded by the $(N-1)$-event analysis, it is not particularly favored either. Our intuition from Section 3, then, suggests that *GW190814's small q is consistent with the rest of the events* (population models already contain plenty of support for small q), but *GW190814 is inconsistent with the rest of the events because of its small* m_2.

We further quantify this by estimating a p-value that the smallest event out of an N-event catalog would have m_2 less than or equal to any event in GWTC-2, given the $(N-1)$-event coarse-grained hyperposterior. We find $P \leqslant 0.056\%$ at 90% confidence.[15] As

[15] We can only bound the p-value from above, as we did not find a single instance where synthetic catalogs generated $\min\{m_2\}$ smaller than GW190814's secondary after ~5000 trials. Similarly, we find $P \leqslant 0.024\%$ with the $(N-1)$-event analysis that neglects coarse graining.

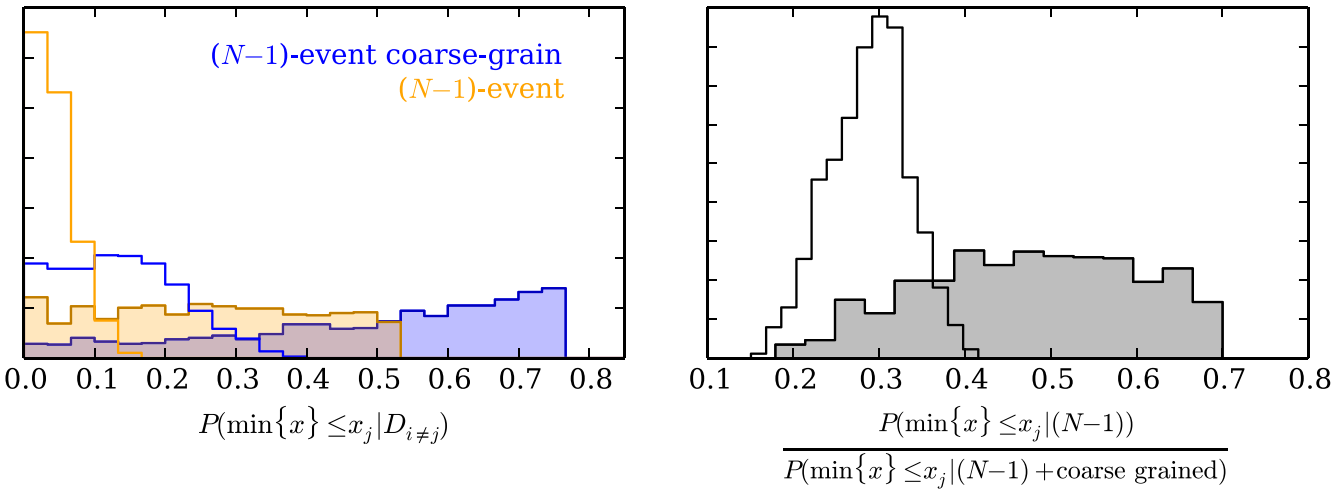

Figure 4. Left: distributions of probabilities that the smallest event in a catalog of 10 events would have $x \leqslant x_j$ (p-values), marginalized over hyperposteriors inferred with the other $N-1$ events when all events are drawn from the same distribution (shaded) and in the presence of a true outlier (unshaded). Colors match Figures 2 and 3. Right: distributions of ratios of p-values from the different $(N-1)$-event hyperposteriors when the null hypothesis is (shaded) true or (unshaded) false. These differences are typically no larger than a factor of a few, although they are more pronounced when the null hypothesis is false. We also note that our p-values are only expected to be uniformly distributed when $\Lambda = \Lambda_0$ (see, e.g., Seth et al. 2019). Because we marginalize over our uncertainty in Λ, the resulting statistic need not be uniformly distributed even when the null hypothesis is true.

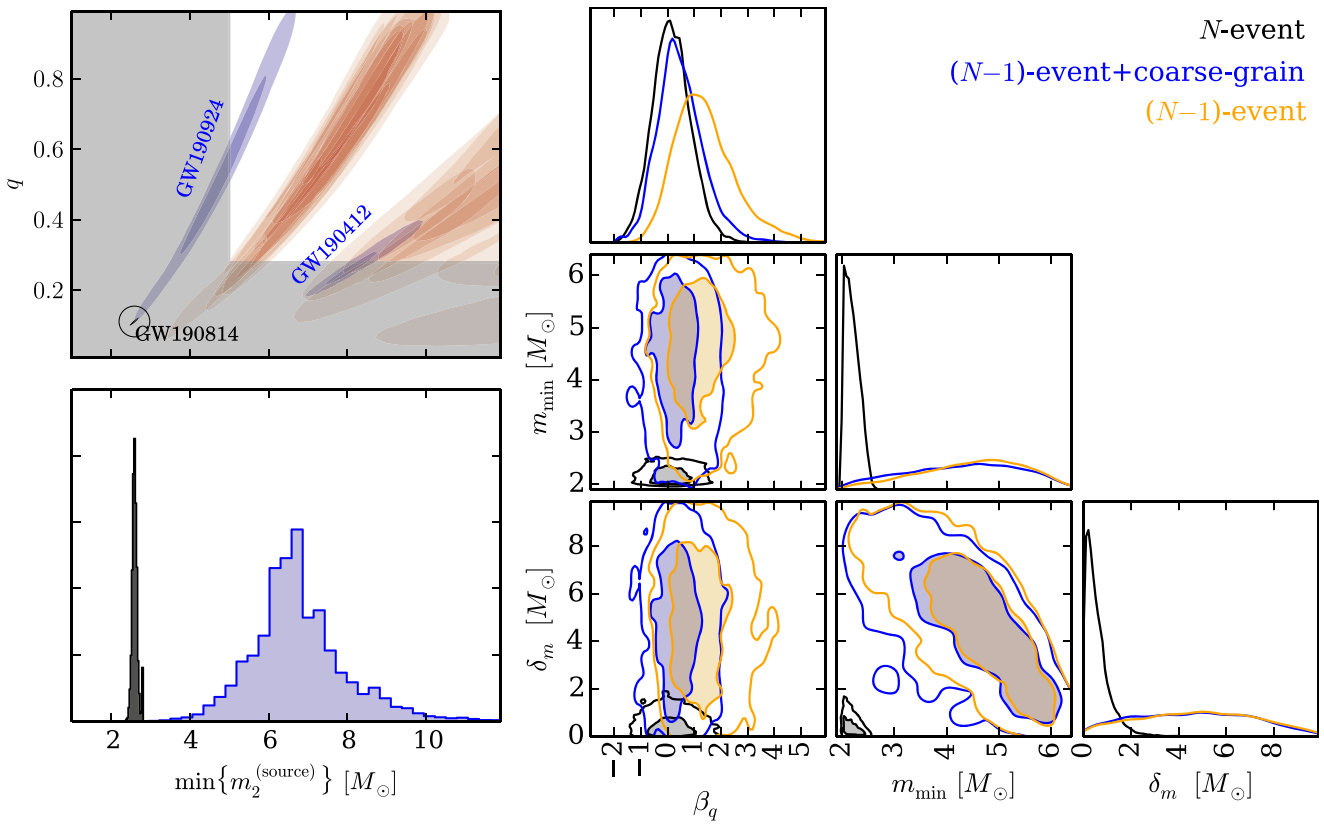

Figure 5. Top left: depiction of $\mathcal{S}_{\text{GW190814}}$ (shaded region, Equation (21)), chosen to be the region ($m_2 \leqslant 5.0\,M_\odot$) *or* ($q \leqslant 0.28$) based on the approximate median m_2 and q of GW190924 and GW190412, respectively. Colored shaded regions show 50% and 90% highest probability density credible regions from each event assuming flat priors in component masses and uniform priors in comoving volume for GW190814 (black ellipse, circled in lower left corner), GW190924 and GW190412 (blue), and all other BBH systems considered in Abbott et al. (2021a; red). Bottom left: distribution of GW190814's m_2 (black) and the smallest expected m_2 (blue) in catalogs of 45 events based on the $(N-1)$-event coarse-grained hyperposterior for the POWERLAW+PEAK mass model. Right: hyperposteriors inferred with different amounts of information about GW190814. Contours in the joint distributions denote 50% and 90% credible regions.

such, we reject the null hypothesis that GW190814 was drawn from the same population as the other $N-1$ events. Abbott et al. (2021a) reach the same conclusion, and, following their example, we exclude GW190814 from the catalog as we explore whether other individual events are inconsistent with the remaining detections.

4.2. GW190412 Is Not an Outlier

We next investigate GW190412 (Abbott et al. 2020c; The LIGO Scientific Collaboration & The Virgo Collaboration 2020c), which, with $q \sim 0.28$, is the only event in GWTC-2 besides GW190814 that is inconsistent with $q = 1$. As a reminder, we have already removed GW190814 from the set of events against which we compare GW190412. In this case, we define

$$\mathcal{S}_{\rm GW190412}\colon\ q \leqslant 0.8 \tag{22}$$

as a conservative boundary for systems that have asymmetric masses. Fishbach & Holz (2020a) estimate that 90% (99%) of detected BBHs will have $q \geqslant 0.73$ (0.51) based on population models of the 10 BBH events in GWTC-1 (Abbott et al. 2019b). Our choice for $\mathcal{S}_{\rm GW190412}$ is even more conservative: we give the coarse-grained inference very little information about GW190412 itself and will therefore most easily identify it as an outlier. Figure 6 demonstrates the results.

Again, we are primarily interested in the low-mass (and low mass ratio) behavior of the model and focus on the inferred values of β_q, $m_{\rm min}$, and δ_m. In this case, we find that all inferences agree remarkably well for $m_{\rm min}$ and δ_m, which is unsurprising since GW190412's masses are not particularly extreme when considered individually. However, we observe better agreement between the N-event and $(N-1)$-event coarse-grained hyperposteriors for β_q than between either of those hyperposteriors and the $(N-1)$-event inference that neglects coarse graining. This is analogous to Figure 3 when the null hypothesis is true; excluding the event with smallest q shifts the inferred hyperposterior toward values that favor equal-mass systems (larger β_q).

The coarse-grained inference produces a p-value of $P = 22\%$[16] for the smallest observed q to be as small as or smaller than that of any BBH event in GWTC-2 (excluding GW190814). We therefore conclude that *GW190412 is consistent with the population inferred from rest of the BBH events within GWTC-2* (except GW190814). It is simply the event with the most extreme q from that population, in agreement with Abbott et al. (2020c, 2021a).

4.3. GW190521 Is Not an Outlier

Finally, we consider GW190521 (Abbott et al. 2020b, 2020d; The LIGO Scientific Collaboration & The Virgo Collaboration 2020d). This event is remarkable for its large component masses, although we note that it does not unambiguously have the largest component masses of any system in GWTC-2; see Figure 7. In this case, we cannot easily define $\mathcal{S}_{\rm GW190521}$ in terms of the other events in the catalog while guaranteeing that Equation (6) is satisfied. However, we note that GW190521 is of particular interest because its component masses nominally fall within the PISN mass gap.[17] In this case, it is natural to define $\mathcal{S}_{\rm GW190521}$ in terms of an approximate boundary defining the PISN mass gap. We take

$$\mathcal{S}_{\rm GW190521}\colon\ m_1 \geqslant 50\,M_\odot \tag{23}$$

as a reasonable approximation, although others have considered values as large as $65\,M_\odot$ (Abbott et al. 2020b, 2020d; O'Brien et al. 2021). Again, there is nontrivial overlap between this choice for $\mathcal{S}_{\rm GW190521}$ and the parameters inferred for several other events in GWTC-2, but this does not affect our inference.

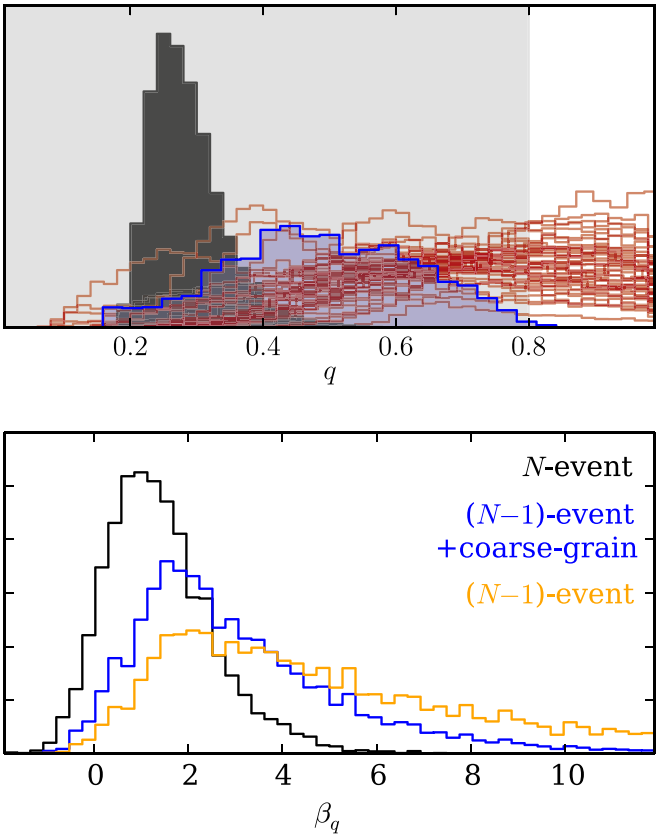

Figure 6. Top: depiction of $\mathcal{S}_{\rm GW190412}$ (shaded gray; Equation (22)) with GW190412 (shaded black) and all other events considered in Abbott et al. (2021a; except GW190814) with flat priors on component masses (red), as well as the prediction for the smallest q out of 44 events based on the $(N-1)$-event coarse-grained hyperposterior with the POWERLAW+PEAK mass model (blue). Bottom: hyperposteriors inferred with different amounts of information about GW190412.

Contrary to GW190814 and GW190412, in which higher-order modes were observed and their asymmetric mass ratios were relatively well measured, GW190521's component mass posterior is quite broad. As such, simultaneous population inference can have a significant impact on the system's inferred properties. For this reason, Figure 7 shows the posteriors of primary masses under the POWERLAW+PEAK mass model as inferred with our coarse-grained analysis. Nonetheless, several analyses have shown that it is likely that at least one component of GW190521 had a mass $\geqslant 65\,M_\odot$ (Abbott et al. 2020b, 2020d; O'Brien et al. 2021).

Some authors have taken this to mean that GW190521 is incompatible with the population of stellar-mass black holes, but, to be clear they often mean the inferred phenomenological fit from other events with an additional cutoff (not included in the fit). We are concerned with the simpler question of whether GW190521 is inconsistent with the phenomenological population model inferred from the other BBH events without imposing additional sharp cutoffs. We are not concerned with whether the current phenomenological models are compatible with PISN predictions, but instead ask whether the models are sufficient to describe the overall mass distribution and whether GW190521 is consistent with models inferred from the other events.

While our coarse-grained hyperposterior at times more closely resembles the $(N-1)$-event hyperposterior than the N-event hyperposterior (bottom panel of Figure 7), we nonetheless obtain

[16] The $(N-1)$-event analysis without coarse graining yields $P = 19\%$.

[17] See, e.g., Fishbach & Holz (2020b) for alternative interpretations with component masses that straddle the mass gap.

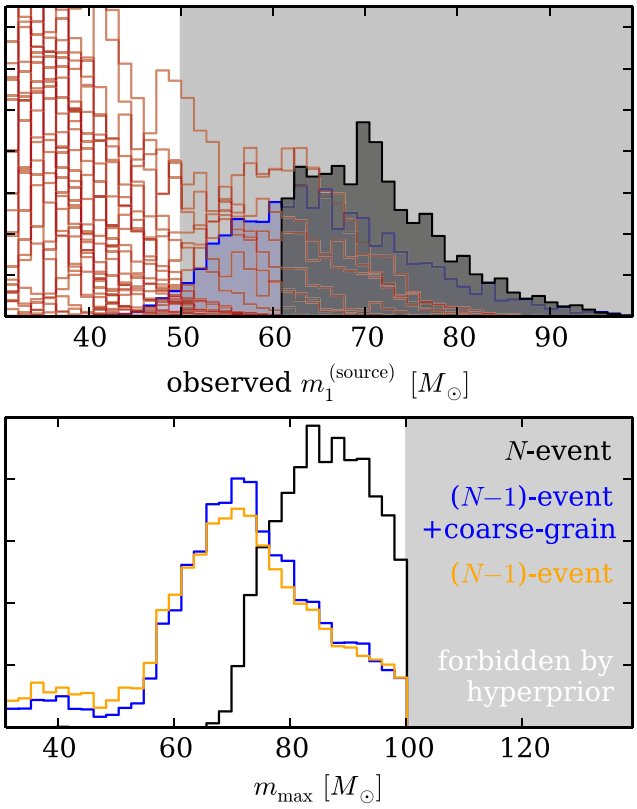

Figure 7. Top: depiction of $\mathcal{S}_{\mathrm{GW190521}}$ (shaded gray; Equation (23)) with GW190521 (black shaded) and all other events considered in Abbott et al. (2021a; red) informed by the POWERLAW+PEAK mass model as inferred under the $(N-1)$-event coarse-grained hyperposterior, as well as the distribution of the largest predicted m_1 out of 44 events based on the $(N-1)$-event coarse-grained hyperposterior. The hard boundary in GW190521's distribution at $m_1 \sim 60\,M_\odot$ is due to our labeling convention ($m_1 \geqslant m_2$) and approximately corresponds to equal-mass systems. Bottom: hyperposteriors inferred with different amounts of information about GW190521. Note that m_{max} extends below $50\,M_\odot$ because it only limits the power-law part of POWERLAW+PEAK; the peak still has support at higher masses.

a p-value of $P = 20\%$ that synthetic catalogs would contain a more massive event than has been observed so far. In fact, if we neglect the coarse-graining correction, we still obtain $P = 15\%$ under the POWERLAW+PEAK model. As such, and in agreement with Abbott et al. (2021a), we conclude that GW190521 is consistent with the overall mass distribution inferred from the rest of the BBH population.

One might expect other mass models, like the Abbott et al. (2021a) TRUNCATED model with a sharp cutoff at high masses, to be less consistent with GW190521. To investigate this, we repeat our coarse-grained analysis and find $P = 11\%$ under the TRUNCATED model. Indeed, even the $(N-1)$-event TRUNCATED hyperposterior that neglects coarse graining only yields $P = 6\%$. As such, we may conclude that, while GW190521 is certainly not expected to be common, it also is not unambiguously inconsistent with any of the phenomenological models considered in Abbott et al. (2021a), even the simplest TRUNCATED distribution.[18]

[18] While GW190521 may not be inconsistent with the mass models considered in Abbott et al. (2021a), they point out that the simple TRUNCATED model is more broadly a bad fit to the data. It overpredicts the number of massive systems that should have been observed (see also Fishbach et al. 2021). However, even minor modifications like the POWERLAW+PEAK and BROKEN POWERLAW distributions appear to remove such tensions.

That being said, the number of BBH systems detected in GWTC-2 more than quadrupled compared to GWTC-1, and both Abbott et al. (2021a) and Fishbach & Holz (2020b) point out that GW190521 does seem to be inconsistent with the truncated mass distributions inferred based on only the 10 events in GWTC-1 (Abbott et al. 2019a). While additional GW detections have continued to surprise, we are reminded that it is important to consider new events in the context of the full catalog before drawing conclusions based on individual (apparently) exceptional events and a subset of previously observed systems.

5. Discussion

Within any set of observed events drawn from an unknown population, one is often interested in determining whether new events are consistent with the population inferred from the existing set. This often involves careful examination of particular events because they are extremal in some way. It is remarkable that GW astronomy has already advanced to the point where such matters are of practical importance in only a half-dozen years since the first detection (Abbott et al. 2016). Nonetheless, we show that current approaches to answer exactly this question, which have become commonplace within the GW community, can introduce biases, as they do not account for the manner in which extremal events were identified for further study.

Our method allows analysts to explicitly account for how they selected extremal events within leave-one-out analyses, representing excluded events with coarse-grained likelihoods, and clearly identifying the need to select the size and placement of the coarse grains. While we note that the exact choice of how big to make those grains is to some degree arbitrary, just as the definition of a null hypothesis is to some degree arbitrary, we propose algorithmic ways to choose the most generous grains possible. We further observe that the resulting coarse-grained analysis almost always has nearly correct coverage within several toy models.

Finally, we note that the biases introduced by excluding extremal events without accounting for how they were selected can be particularly severe when there are sharp features in the underlying population model (e.g., mass gaps). Therefore, one must take care when analyzing population models with sharp cutoffs and attempting to assess the significance of outliers after excluding extremal events. However, even in these severe cases, we find that p-values estimated from $(N-1)$-event hyperposteriors are typically biased by at most a factor of a few.

Our conclusions based on our coarse-grained analysis agree with those presented in Abbott et al. (2021a), even though they did not account for the coarse-grained correction and their leave-one-out analyses may have been biased. We find that *GW190814 is an outlier* because its secondary mass is too small to be consistent with the other events. *GW190412 is not an outlier*, and its small mass ratio is simply the most extreme example from the tail of the main population. We find that *GW190521 is not an outlier* under the preferred mass models explored in Abbott et al. (2021a) and is in only moderate tension with even the simplest truncated mass models.

We again note that any population analysis will eventually face the challenge of determining whether particular events are consistent with the population inferred from the rest of the events. This problem is not unique to GW astronomy. However, as catalogs continue to rapidly grow in size, this question has become increasingly relevant. We note that another large set of events is expected with the release of the second half of the LVK collaborations' third observing run (O3b). Indeed, given the

breadth of physical phenomena that GW observations can probe, it is of the utmost importance to fully characterize outlier tests. We emphasize that many of the most interesting questions in GW astronomy are specifically focused on outliers, including extreme mass, mass ratio, and spin events, and the presence of events within the putative NS–BH and PISN mass gaps. Our analysis provides a controlled way to account for event selection when examining outliers with nearly trivial additional computational cost. This will enable the robust identification of novel subpopulations without fear of biasing analyses toward artificially inflated significance estimates for potential outliers.

The authors thank Will Farr, Tom Callister, and Katerina Chatziioannou for several helpful discussions. R.E. thanks the Canadian Institute for Advanced Research (CIFAR) for support. Research at Perimeter Institute is supported in part by the Government of Canada through the Department of Innovation, Science and Economic Development Canada and by the Province of Ontario through the Ministry of Colleges and Universities. S.G. and E.T. are supported through Australian Research Council (ARC) Centre of Excellence CE170100004. A.F. is supported by the NSF Research Traineeship program under grant DGE-1735359. M.F. is supported by NASA through NASA Hubble Fellowship grant HST-HF2-51455.001-A awarded by the Space Telescope Science Institute. D.E.H. is supported by NSF grants PHY-2006645, PHY-2011997, 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 on work supported by NSF LIGO Laboratory, which is a major facility fully funded by the National Science Foundation. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation grants PHY-0757058 and PHY-0823459.

Appendix A
Reweighing Existing Hyperposteriors

We note that Equation (10) could be computationally expensive if we allow $\mathcal{S}$ to depend on the parameters of the other events. However, if $\mathcal{S}$ does not depend on the parameters of the other events, the marginalization becomes trivial. There are also significant redundancies between Equations (5) and (10), which reduce the computational cost of reweighing existing samples and make implementing the coarse-grained likelihood a simple extension of existing likelihoods.

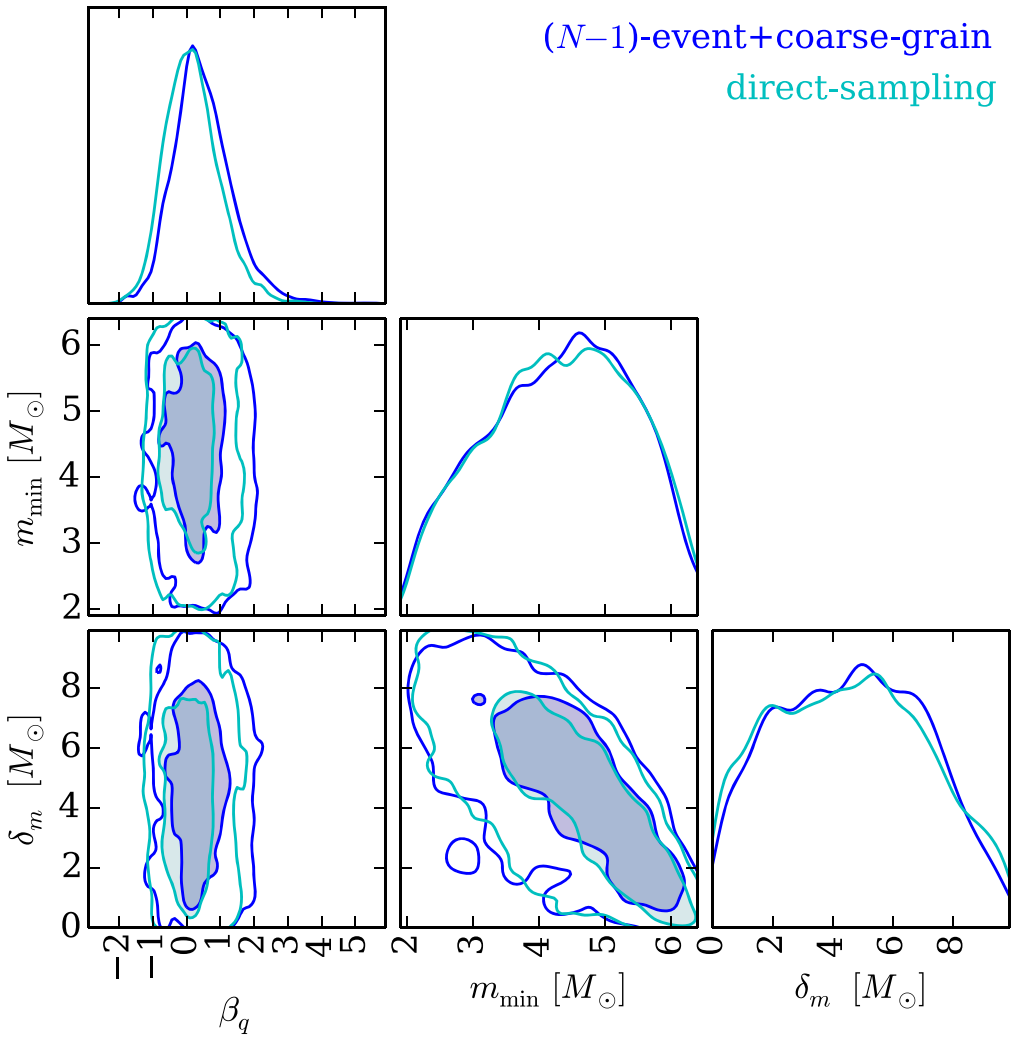

Figure 8. Comparison between reweighed samples and direct sampling for GW190814's $(N-1)$+coarse-grained hyperposterior (Section 4.1). Contours in the joint distributions denote 50% and 90% credible regions. Reweighed samples were obtained by applying Equation (A2) to public hyperposterior samples from an $(N-1)$-event analysis that excluded GW190814 (The LIGO Scientific Collaboration & The Virgo Collaboration 2020a).

There are two corrections to the likelihood that may be of interest. When we have already analyzed the full set of N events and want to conduct a coarse-grained analysis post hoc, we note that

$$\begin{aligned}\frac{p(\Lambda|\{D_{i\neq j}\},\, \rho(D_j) \geqslant \rho_{\rm thr};\, \theta_j \in \mathcal{S})}{p(\Lambda|\{D_i\})} &\propto \frac{\int d\theta\, p(\theta|\Lambda) P(\det|\theta)\Theta(\theta \in \mathcal{S}(\{\theta_{i\neq j}\}))}{\int d\theta_j\, p(D_j|\theta_j) p(\theta_j|\Lambda)} \\ &= \frac{P(\theta \in \mathcal{S},\, \det|\Lambda)}{p(D_j|\Lambda)}. \end{aligned} \tag{A1}$$

When we instead have hyperposterior samples from an $(N-1)$-event analysis that omitted the potential outlier, we can write

$$\begin{aligned}\frac{p(\Lambda|\{D_{i\neq j}\},\, \rho(D_j) \geqslant \rho_{\rm thr};\, \theta_j \in \mathcal{S})}{p(\Lambda|\{D_{i\neq j}\})} &\propto P(\theta \in \mathcal{S}|\,\det;\Lambda). \end{aligned} \tag{A2}$$

Weighing existing hyperposterior samples by either Equation (A1) or Equation (A2), as appropriate, allows us to estimate the coarse-grained hyperposterior and quickly perform consistency tests of our null hypothesis. Such reweighing procedures are equivalent to directly sampling from Equation (10) as long as enough effective samples remain to provide reliable estimates of the hyperposterior. As a demonstration, we repeat the analysis of Section 4.1 by directly sampling from Equation (10), obtaining equivalent results to the reweighed hyperposterior samples (see Figure 8).

Appendix B
Toy Models with Larger Catalogs

Section 3 examines the effect of coarse graining on catalogs with 10 events in the context of a simple toy model. Here we investigate how our conclusions might change for catalogs of different sizes. Specifically, we consider three cases:

1. Null hypothesis is true, we only have "regular" events, but we have more of them (Appendix B.1).
2. Null hypothesis is false, we have a single "true outlier," but we have more regular events (Appendix B.2).
3. Null hypothesis is false, we have more events, but we retain an equal ratio of true outliers and regular events (Appendix B.3).

In each case, we compare how our conclusions behave when we exclude events based on x or based on $q \equiv x/y$.

B.1. Null Hypothesis Is True

We begin by analyzing the relative importance of coarse graining when the null hypothesis is true (all events are drawn from the same distribution). Figure 9 demonstrates our results by analyzing how the $(N-1)$-event hyperposterior's coverage changes as we increase the number of events in our catalog (N). Because we believe that the N-event and $(N-1)$-event + coarse-grain hyperposteriors have correct coverage, we focus only on the $(N-1)$-event hyperposterior that neglects the coarse-graining correction.

When we exclude the event with the smallest x, we see very similar coverage regardless of the number of events in the catalog. That is, while the $(N-1)$ hyperposteriors become more precise and the smallest observed x approaches m, this occurs in such a way that the coverage is almost unchanged. If anything, the coverage appears to become slightly worse as N increases. The $(N-1)$-event hyperposterior is consistently shifted relative to the N-event hyperposterior. This is due to the fact that the sharp edge within the population model asserts that we know the exact shape of the distribution very precisely.

Therefore, in the presence of a sharp feature in the population model, we might expect to get a hyperposterior that is closer to the truth (more accurate and precise), but it will still consistently assign the true hyperparameters posterior probabilities that are too low (ratio of accuracy to precision is approximately the same).

We see different behavior when we exclude the event with the smallest q. In this case, the $(N-1)$-event analysis that neglects the coarse-grain correction sees less of a bias as N increases. That is, the coverage approaches a diagonal line to within the resolution of the finite number of trials performed and the $(N-1)$-event hyperposterior collapses to the same result we obtain for the N-event hyperposterior. We attribute this to the fact that there is not a sharp feature in our population model for q. Therefore, the fact that we incorrectly exclude a single event does not add an undue amount of information to the analysis, and the larger number of overall events tends to overwhelm the impact of the single excluded event.

We therefore conclude that coarse graining could be important regardless of the catalog's size, but only if one assumes a model with a sharp feature in the dimension used to define $\mathcal{S}$.

B.2. Null Hypothesis Is False with a Single True Outlier

Next, we consider the case where the null hypothesis is false, but there is always only a single true outlier. As in Figures 2 and 3, we find good agreement between both $(N-1)$-event hyperposteriors, and both show good coverage for almost all catalog sizes. Nevertheless, the $(N-1)$-event analysis that neglects coarse graining always predicts smaller p-values for the smallest observed event than the $(N-1)$-event analysis that includes coarse graining.

For both $(N-1)$-event analyses, the p-values decrease as the catalog size increases. This means we are able to more confidently identify true outliers (when only a single outlier exists) within larger catalogs, as expected. We therefore focus on the ratio of p-values as a way to evaluate the relative sensitivity of each analysis.

Figure 10 shows that this ratio tends to 1 when we exclude the event with the smallest q, suggesting that both analyses are equally sensitive to true outliers within large catalogs when there are no sharp features in the population model. Conversely, we observe an approximately stationary distribution when we exclude the smallest observed x, regardless of catalog size. We can again understand this from the fact that the hyperposteriors tend toward one another (relative to their widths) as N increases when we exclude the smallest q, but they remain consistently shifted when we exclude the smallest x. Therefore, as is the case when the null hypothesis is true, coarse graining may always be important regardless of the catalog's size when there is a sharp feature in the population model.

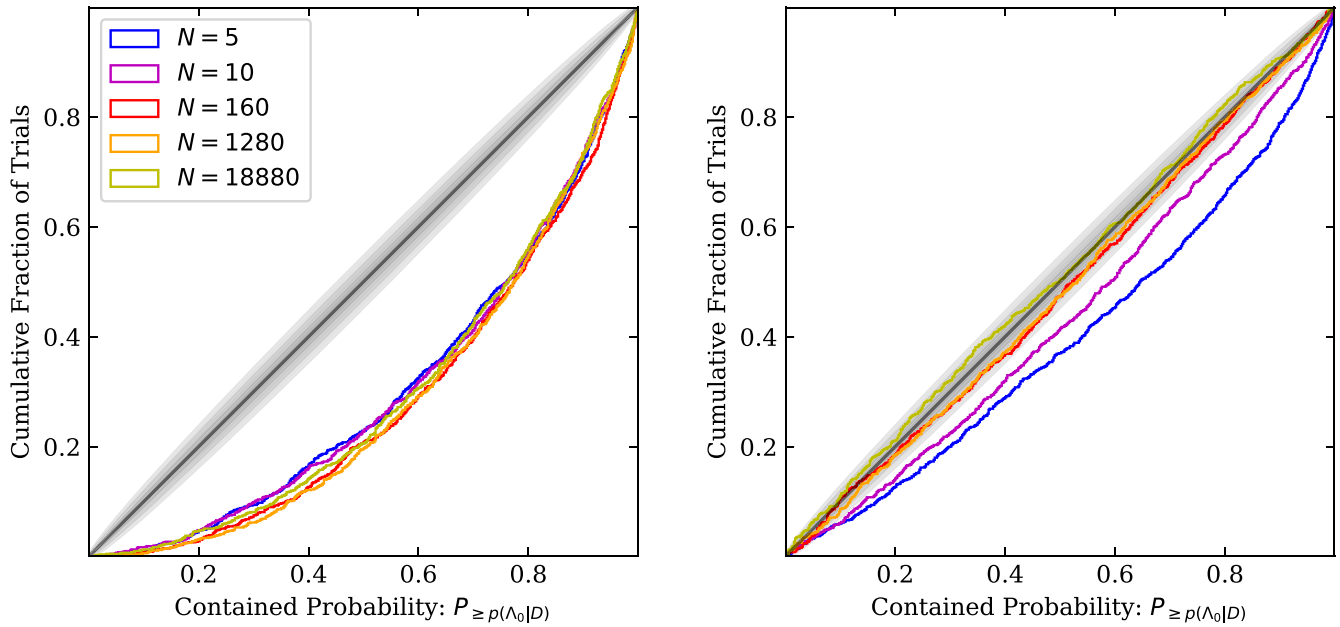

Figure 9. Coverage plots showing how the $(N-1)$ hyperposterior that neglects coarse graining behaves as we increase the number of events in the catalog (N) over 3 orders of magnitude when we exclude the event with the smallest x (left) and q (right).

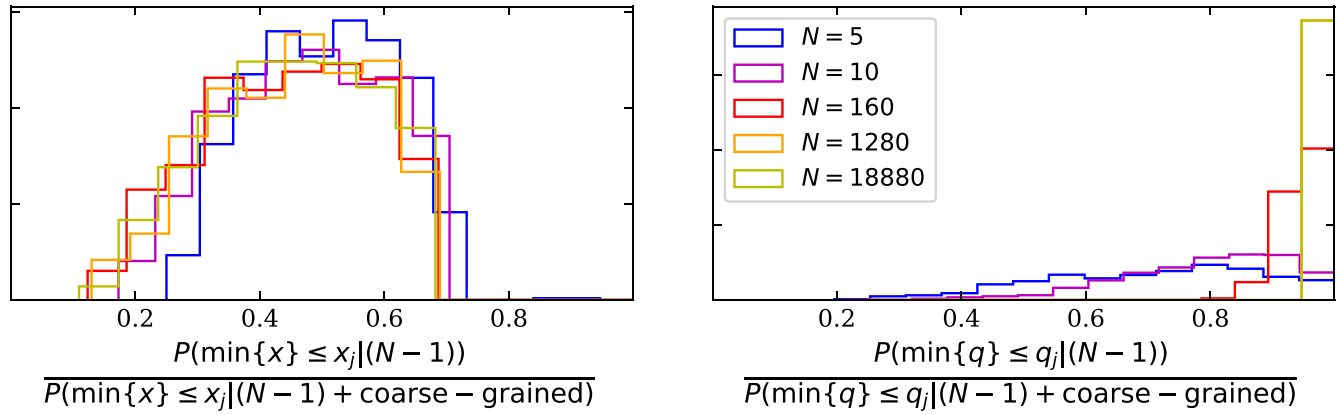

Figure 10. Ratios of p-values between the $(N-1)$-event analyses when the null hypothesis is false and there is a single true outlier, analogous to the unshaded distribution in the right panel of Figure 4. We show how the ratio changes when we remove the event with the smallest observed x (left) and q (right). In all cases, the p-values themselves become vanishingly small, and we can confidently identify the true outlier.

B.3. Null Hypothesis Is False with an Equal Ratio of Outliers and Regular Events

Finally, we consider the case where the number of "true outliers" scales with the size of the catalog. This situation is mostly likely to occur for real astrophysical sources. As we observe for longer periods, we expect to see the same proportion of events from each subpopulation. From this perspective, a single outlier (Appendix B.2) that is not repeated as the catalog grows may suggest that it is not of astrophysical origin.

We expect that the relative importance of coarse graining will increase when we observe more events from all subpopulations. That is, if the population model describing regular events does not support a subpopulation, we expect to need to coarse-grain all events in the subpopulation to obtain a robust inference of the main population. In this case, we might expect comparable shifts in the hyperposteriors as are seen in Section 3. At the same time, the statistical uncertainty on that inference should improve, meaning that the shifts observed in the hyperposteriors will be more statistically significant. Based on our previous examples, this may be more important for population models with sharp features. However, it may be of more practical use to modify the population model to include the apparent subpopulation rather than coarse-graining the entire subpopulation.

ORCID iDs

Reed Essick https://orcid.org/0000-0001-8196-9267
Amanda Farah https://orcid.org/0000-0002-6121-0285
Shanika Galaudage https://orcid.org/0000-0002-1819-0215
Colm Talbot https://orcid.org/0000-0003-2053-5582
Maya Fishbach https://orcid.org/0000-0002-1980-5293
Eric Thrane https://orcid.org/0000-0002-4418-3895
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. 2016, PhRvL, 116, 061102

Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2019a, ApJL, 882, L24
Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2019b, PhRvX, 9, 031040
Abbott, R., Abbott, T. D., Abraham, S., et al. 2020a, ApJL, 896, L44
Abbott, R., Abbott, T. D., Abraham, S., et al. 2020b, ApJL, 900, L13
Abbott, R., Abbott, T. D., Abraham, S., et al. 2020c, PhRvD, 102, 043015
Abbott, R., Abbott, T. D., Abraham, S., et al. 2020d, 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. 2021b, PhRvX, 11, 021053
Acernese, F., Agathos, M., Agatsuma, K., et al. 2014, CQGra, 32, 024001
Baxter, E. J., Croon, D., McDermott, S. D., & Sakstein, J. 2021, ApJL, 916, L16
Belczynski, K., Wiktorowicz, G., Fryer, C. L., Holz, D. E., & Kalogera, V. 2012, ApJ, 757, 91
Croon, D., McDermott, S. D., & Sakstein, J. 2020, PhRvD, 102, 115024
De Luca, V., Desjacques, V., Franciolini, G., Pani, P., & Riotto, A. 2021, PhRvL, 126, 051101
Edelman, B., Doctor, Z., & Farr, B. 2021, ApJL, 913, L23
Essick, R., & Landry, P. 2020, ApJ, 904, 80
Ezquiaga, J. M., & Holz, D. E. 2021, ApJL, 909, L23
Farmer, R., Renzo, M., de Mink, S. E., Fishbach, M., & Justham, S. 2020, ApJL, 902, L36
Farr, W. M., Fishbach, M., Ye, J., & Holz, D. E. 2019, ApJL, 883, L42
Fishbach, M., Doctor, Z., Callister, T., et al. 2021, ApJ, 912, 98
Fishbach, M., Farr, W. M., & Holz, D. E. 2020, 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
Franciolini, G., Baibhav, V., De Luca, V., et al. 2021, arXiv:2105.03349
Fryer, C. L., & Kalogera, V. 2001, ApJ, 554, 548
Gerosa, D., & Fishbach, M. 2021, NatAs, 5, 749
Heger, A., & Woosley, S. E. 2002, ApJ, 567, 532
Kimball, C., Talbot, C., Berry, C. P. L., et al. 2021, ApJL, 915, L35
The LIGO Scientific Collaboration & The Virgo Collaboration 2020a, Data Release for Population Properties of Compact Objects from the Second LIGO-Virgo Gravitational-Wave Transient Catalog, https://dcc.ligo.org/LIGO-P2000434/public
The LIGO Scientific Collaboration & The Virgo Collaboration 2020b, GW190814 Data Release, https://www.gw-openscience.org/eventapi/html/O3_Discovery_Papers/GW190814/v1/
The LIGO Scientific Collaboration & The Virgo Collaboration 2020c, GW190521 Data Release, https://www.gw-openscience.org/eventapi/html/O3_Discovery_Papers/GW190412/v1/
The LIGO Scientific Collaboration & The Virgo Collaboration 2020d, GW190521 Data Release, https://www.gw-openscience.org/eventapi/html/O3_Discovery_Papers/GW190521/v2/
Mandel, I., Farr, W. M., Colonna, A., et al. 2016, MNRAS, 465, 3254
Miller, J. W., & Dunson, D. B. 2019, J. Am. Stat. Assoc., 114, 1113
Ng, K. K. Y., Vitale, S., Farr, W. M., & Rodriguez, C. L. 2021, ApJL, 913, L5
Nitz, A. H., & Capano, C. D. 2021, ApJL, 907, L9
O'Brien, B., Szczepanczyk, M., Gayathri, V., et al. 2021, PhRvD, 104, 082003
Seth, S., Murray, I., & Williams, C. K. I. 2019, BayAn, 14, 703
Tagawa, H., Kocsis, B., Haiman, Z., et al. 2021, ApJ, 908, 194
Talbot, C., & Thrane, E. 2018, ApJ, 856, 173
Thomas, O., & Corander, J. 2019, arXiv:1912.05810
Thrane, E., & Talbot, C. 2019, PASA, 36, E010
Vehtari, A., Gelman, A., & Gabry, J. 2017, Stat. Comput., 27, 1413
Vitale, S., Gerosa, D., Farr, W. M., & Taylor, S. R. 2020, arXiv:2007.05579
Zevin, M., Bavera, S. S., Berry, C. P. L., et al. 2021, ApJ, 910, 152
Zevin, M., Spera, M., Berry, C. P. L., & Kalogera, V. 2020, ApJL, 899, L1