The Astrophysical Journal, 979:9 (13pp), 2025 January 20 https://doi.org/10.3847/1538-4357/ad9393

Gravitational-wave Dark Siren Cosmology Systematics from Galaxy Weighting

Alexandra G. Hanselman[1], Aditya Vijaykumar[1,2], Maya Fishbach[2,3], and Daniel E. Holz[1,4,5,6]

[1] Department of Physics, The University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA; aghanselman@uchicago.edu, aditya@utoronto.ca
[2] Canadian Institute for Theoretical Astrophysics, University of Toronto, 60 St George St, Toronto, ON M5S 3H8, Canada
[3] David A. Dunlap Department of Astronomy and Astrophysics, and Department of Physics, 60 St George St, University of Toronto, Toronto, ON M5S 3H8, Canada
[4] Enrico Fermi Institute, The University of Chicago, 933 East 56th Street, Chicago, IL 60637, USA
[5] Department of Astronomy and Astrophysics, The University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA
[6] Kavli Institute for Cosmological Physics, The University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA

Received 2024 June 19; revised 2024 November 10; accepted 2024 November 13; published 2025 January 13

Abstract

The detection of GW170817 and the measurement of its redshift from the associated electromagnetic counterpart provided the first gravitational-wave (GW) determination of the Hubble constant (H_0), demonstrating the potential power of standard siren cosmology. In contrast to this “bright siren” approach, the “dark siren” approach can be utilized for GW sources in the absence of an electromagnetic counterpart: One considers all galaxies contained within the localization volume as potential hosts. When statistically averaging over the potential host galaxies, weighting them by physically motivated properties (e.g., tracing star formation or stellar mass) could improve convergence. Using mock galaxy catalogs, we explore the impact of these weightings on the measurement of H_0. We find that incorrect weighting schemes can lead to significant biases due to two effects: the assumption of an incorrect galaxy redshift distribution, and preferentially weighting incorrect host galaxies during the inference. The magnitudes of these biases are influenced by the number of galaxies along each line of sight, the measurement uncertainty in the GW luminosity distance, and correlations in the parameter space of galaxies. We show that the bias may be overcome from improved localization constraints in future GW detectors, a strategic choice of priors or weighting prescription, and by restricting the analysis to a subset of high-signal-to-noise ratio events. We propose the use of hierarchical inference as a diagnostic of incorrectly weighted prescriptions. Such approaches can simultaneously infer the correct weighting scheme and the values of the cosmological parameters, thereby mitigating the bias in dark siren cosmology due to incorrect host-galaxy weighting.

Unified Astronomy Thesaurus concepts: Gravitational wave astronomy (675); Cosmology (343); Hubble constant (758); Gravitational waves (678); Gravitational wave sources (677)

1. Introduction

Measuring the expansion rate of the Universe has been a key goal of observational cosmology for almost a century. Specifically, the local expansion rate of the Universe, the Hubble constant (H_0), has recently been a topic of intense debate. Precision measurements of H_0 from low-redshift probes (e.g., supernovae; D. Scolnic et al. 2022; S. A. Uddin et al. 2023) and high-redshift probes (e.g., cosmic microwave background; Planck Collaboration et al. 2020) disagree, giving rise to the “H_0 tension” (see E. Di Valentino et al. 2021; W. L. Freedman & B. F. Madore 2023). Alternate observational probes of H_0 are of particular utility in distinguishing whether the tension is due to unmodeled systematics in current observations or physics beyond the standard model of cosmology.

Observations of gravitational waves (GWs) from compact binary coalescences (CBCs) have been proposed as probes of cosmic expansion (B. F. Schutz 1986; D. E. Holz & S. A. Hughes 2005). The luminosity distance to a CBC can be estimated directly from the gravitational waveform (B. F. Schutz 1986; L. S. Finn & D. F. Chernoff 1993; C. Cutler & É. Flanagan 1994) without reference to a distance ladder. If the cosmological redshift of the CBC can be estimated by some other means, it is possible to infer the values of H_0 and other cosmological parameters governing the expansion history of the Universe. For instance, binary neutron stars have electromagnetic counterparts which can enable the localization of the source to its host galaxy, yielding a redshift measurement (D. E. Holz & S. A. Hughes 2005; N. Dalal et al. 2006; S. Nissanke et al. 2010, 2013; H.-Y. Chen et al. 2018). The observation of an electromagnetic counterpart from GW170817 led to the identification of NGC 4993 as the source’s host galaxy (B. P. Abbott et al. 2017a, 2017b, 2017c; D. A. Coulter et al. 2017; M. Soares-Santos et al. 2017), yielding a ∼15% measurement of H_0 solely from this source (B. P. Abbott et al. 2017d). However, subsequent observing runs of the LIGO-Virgo-KAGRA (LVK) collaboration have to date failed to yield additional mergers with electromagnetic counterparts (R. Abbott et al. 2023a).

In the absence of electromagnetic counterparts, Bernard Schutz proposed an alternative “dark siren” method where one considers all galaxies in the localization volume of a given CBC as potential hosts (B. F. Schutz 1986; C. L. MacLeod & C. J. Hogan 2008; W. Del Pozzo 2012; H.-Y. Chen et al. 2018; M. Fishbach et al. 2019; M. Soares-Santos et al. 2019; R. Gray et al. 2020; A. Palmese et al. 2020; B. P. Abbott et al. 2021; A. Finke et al. 2021; M. Mancarella et al. 2022; R. Abbott et al. 2023b; A. Palmese et al. 2023). For a typical CBC, the number of galaxies in a typical localization volume is large (e.g., ∼408 for GW170817; M. Fishbach et al. 2019); consequently, the H_0 measurement from a single event using this method is generally highly uncertain. However, the expectation is that H_0

measurements stacked over multiple events would reduce this uncertainty, enabling an H_0 measurement of a few percent. This statistical dark siren technique yields a ∼20% measurement of H_0 from dark sirens alone in current data (M. Mancarella et al. 2022; R. Abbott et al. 2023b; A. Palmese et al. 2023). Other methods of redshift identification have been proposed, including using the large-scale two-point cross-correlation between galaxies and GW mergers (T. Namikawa et al. 2016; S. Bera et al. 2020; S. Mukherjee et al. 2021; Cigarrán Díaz & S. Mukherjee 2022), identification of features in the mass spectrum ("spectral sirens"; S. R. Taylor et al. 2012; W. M. Farr et al. 2019; S. Mastrogiovanni et al. 2021; J. M. Ezquiaga & D. E. Holz 2022), and harnessing information from the equation of state of dense matter ("Love sirens"; C. Messenger & J. Read 2012; D. Chatterjee et al. 2021).

One of the ingredients that goes into the statistical dark siren method is the probability that a particular galaxy is the host of the CBC based on its physical properties (C. Messenger & J. Read 2012; M. Fishbach et al. 2019; R. Gray et al. 2020). For instance, depending on the delay-time distribution of CBCs, they could either preferentially merge in star-forming galaxies (short delay times) or massive galaxies (long delay times; see, e.g., S. Adhikari et al. 2020; A. Vijaykumar et al. 2024).[7] While analyzing events to infer H_0, this information can be folded in by weighting each candidate host galaxy by its luminosity within a certain bandpass that best tracks the desired physical quantities, e.g., a galaxy's B-band luminosity as a proxy for star formation rate or K-band luminosity for its total stellar mass (E. F. Bell et al. 2003; L. P. Singer et al. 2016). Note that choosing to not preferentially weight galaxies based on their physical properties amounts to applying equal weights to all galaxies, and is also an imposition of prior belief about the galaxies that host GWs. Unfortunately, conclusively inferring the host-galaxy distribution from data is difficult, owing to poor sky localization of GW sources. Therefore, diagnosing the impact of incorrect weighting schemes is imperative to understanding any systematics associated with the statistical dark siren approach.

E. Trott & D. Huterer (2023) argue that any results obtained using the dark siren approach would be biased in general. However, J. R. Gair et al. (2023) demonstrate explicitly that their arguments are incorrect. In particular, as long as the dark siren cosmology method is applied consistently, the results are unbiased. However, neither of these works consider potential biases from the weighting of host galaxies. In this work, we explore the impact that an incorrect weighting scheme would have on H_0 inference. We do so by building mock catalogs of GW sources and their host galaxies under physically motivated weights, and explore how the inference is affected by changing the weighting schemes. In general, we find that assuming the correct galaxy host weighting scheme leads to an unbiased H_0 estimate, but assuming an incorrect weighting scheme can lead to substantial biases. We also note here that these systematics are different compared to other systematics that impact dark siren cosmology, e.g., models for the mass distribution of CBCs (G. Pierra et al. 2024) and photometric redshift uncertainties in galaxy surveys (C. Turski et al. 2023). During the final stages of this work, G. Perna et al. (2024) completed a related investigation using the MICECAT mock galaxy catalog (J. Carretero et al. 2015; M. Crocce et al. 2015; P. Fosalba et al. 2015a, 2015b; K. Hoffmann et al. 2015), considering the case where the true GW merger rates follow the galaxy star formation rate, with host galaxies weighted by K-band luminosities. When investigating biases due to mismatch between the GW merger rates and recovered weighting schemes, G. Perna et al. (2024) find broadly consistent conclusions to what we report below. Here, we expand on the different combinations of injection and recovery weighting schemes, investigate the underlying causes for the biases, and propose potential methods to diagnose and mitigate these biases.

The rest of this paper is organized as follows. In Section 2, we describe our inference and data-generation prescription. In Section 3, we identify areas for potential bias. We investigate possible diagnostics and discuss various factors that influence potential biases in Section 4, and finally summarize our results in Section 5.

2. Methods

2.1. Inferring H$_0$ Using a Bayesian Scheme

We use the H_0 inference prescription outlined in J. R. Gair et al. (2023) with some modifications that we summarize below. Let us assume we have a set of $N_{\rm GW}$ GW observations with observed data, $\{\hat{x}\}$, where we take the only important quantity to be the observed luminosity distances $\{\hat{d_L}\}$ to be consistent with J. R. Gair et al. (2023).[8] Using Bayes' theorem, the posterior on H_0 is given by

$$p(H_0|\{\hat{d_L}\}) \propto \mathcal{L}(\{\hat{d_L}\}|H_0)p(H_0), \tag{1}$$

where $p(H_0)$ is the prior on H_0, which we take to be uniform, and $\mathcal{L}(\{\hat{d_L}\}|H_0)$ is the likelihood of observing the data $\{\hat{d_L}\}$ given a value for H_0. The likelihood can be further written as (J. R. Gair et al. 2023)

$$\mathcal{L}(\hat{d_L}^i|H_0) = \frac{\int dz\, \mathcal{L}_{\rm GW}(\hat{d_L}^i|d_L(z, H_0))\, p_{\rm CBC}(z, w)}{\int dz\, P_{\rm det}^{\rm GW}(\hat{d_L}^i|d_L(z, H_0))\, p_{\rm CBC}(z, w)}, \tag{2}$$

where $\mathcal{L}_{\rm GW}$ is the likelihood of measuring an observed luminosity distance $\hat{d_L}^i$ given a galaxy with true luminosity distance $d_L(z, H_0)$, and $P_{\rm det}^{\rm GW}$ is the GW detection probability, where we take a GW to be detected if its measured luminosity distance is positive and less than the detection threshold, which we define as $\hat{d_L}^{\rm thr} = 1550$ Mpc to be consistent with J. R. Gair et al. (2023), unless otherwise specified. We also include an extra term, w, in the likelihood, which allows us to assign weights to galaxies in our catalog (see, e.g., M. Fishbach et al. 2019; R. Gray et al. 2020 and references therein). In this work, we take weights that are conditionally independent of redshift and that account for physically motivated quantities, or use equal weights to consider the case where no preferential probability is given to any galaxy. We provide further details

[7] See also other works that forward model the distribution of GW host galaxies based on population synthesis models (R. O'Shaughnessy et al. 2010; A. Lamberts et al. 2016; M. C. Artale et al. 2019; M. Mapelli et al. 2019; M. Toffano et al. 2019; F. Santoliquido et al. 2022; L. Rauf et al. 2023; R. Srinivasan et al. 2023).

[8] In more realistic analyses, other GW parameters, such as sky probability, are relevant. Since we are only considering this toy model, we do not write out the full likelihood here, but point readers to J. R. Gair et al. (2023) and references therein for a full derivation.

on these weights in Section 2.2. We assume the likelihood $\mathcal{L}_{\rm GW}$ is a Gaussian such that

$$\mathcal{L}_{\rm GW}(\hat{d}_L^{\,i}|d_L(z, H_0)) = \frac{1}{\sqrt{2\pi}Ad_L(z, H_0)}\exp\left[-\frac{1}{2}\frac{(\hat{d}_L^{\,i} - d_L(z, H_0))^2}{(Ad_L(z, H_0))^2}\right], \tag{3}$$

where A is a constant fractional error. This yields a detection probability of

$$\begin{aligned} P_{\rm det}^{\rm GW}(\hat{d}_L|d_L(z_j^{\rm gal}, H_0)) &= \int_{-\infty}^{\infty}\Theta(\hat{d}_L^{\,\rm thr} - \hat{d}_L)\Theta(\hat{d}_L)\mathcal{L}_{\rm GW}(\hat{d}_L|d_L(z_j^{\rm gal}, H_0))d\hat{d}_L \\ &= \frac{1}{2}\left[\mathrm{erf}\left(\frac{1}{\sqrt{2}A}\right) - \mathrm{erf}\left(\frac{d_L(z_j^{\rm gal}, H_0) - \hat{d}_L^{\,\rm thr}}{\sqrt{2}Ad_L(z_j^{\rm gal}, H_0)}\right)\right], \end{aligned} \tag{4}$$

where we add an extra bound on $\hat{d}_L$ to ensure that the measured luminosity distance is always positive. Finally, $p_{\rm CBC}(z, w) = \sum_{j=1}^{N_{\rm GAL}} w_j\delta(z - z_j^{\rm gal})$ is the probability that a CBC is at a given redshift z, which we assume to be a delta function for each galaxy in our catalog, as we take galaxy redshifts to be perfectly known (see Section 2.2), and is weighted by the probability of any galaxy to be the host. In the above prescription, we forgo the $(1+z)^{-1}$ conversion from source to detector frame rate; this will not affect our results since they are consistent in both the simulated data set and the recovery procedure. These assumptions simplify the likelihood to

$$\mathcal{L}(\hat{d}_L^{\,i}|H_0) = \frac{\sum_{j=1}^{N_{\rm GAL}}\mathcal{L}_{\rm GW}(\hat{d}_L^{\,i}|d_L(z_j^{\rm gal}, H_0))w_j}{\sum_{j=1}^{N_{\rm GAL}}P_{\rm det}^{\rm GW}(\hat{d}_L^{\,i}|d_L(z_j^{\rm gal}, H_0))w_j}, \tag{5}$$

with $\mathcal{L}_{\rm GW}$ and $P_{\rm det}^{\rm GW}$ now given by Equations (3) and (4). Note that in the above prescription, we ignore sky positions and instead vary the total number of galaxies in a given line of sight , $N_{\rm GAL}$, as a proxy for varying sky position uncertainty. For reference, an $N_{\rm GAL}$ of 10,000 assuming a galaxy number density of 0.01 Mpc^{-3} would yield a localization volume of approximately 10^6 Mpc^3.

2.2. *Mock Galaxy Catalog*

The mock catalog we use in this work is created following the UniverseMachine semi-analytical galaxy formation simulations (P. Behroozi et al. 2019). UniverseMachine starts off with a pure dark matter simulation and populates galaxies into halos using a Monte Carlo scheme while ensuring that their derived properties (star formation histories, stellar masses, etc.) are consistent with a wide range of observations. For our purposes, UniverseMachine provides a distribution of galaxies with physical properties such as stellar mass (SM) and star formation rate (SFR) for bins in redshift.[9] In this work, we take our physically motivated weights to be either the galaxy stellar mass or star formation rate, e.g., $w_i \propto \{\mathrm{SM}_i, \mathrm{SFR}_i\}$. These are common choices, although in more realistic analyses the typically chosen weights are a galaxy's K-band or B-band luminosity as a proxy for either SM or SFR, respectively (see, e.g., M. Fishbach et al. 2019; R. Abbott et al. 2023b).

Throughout this paper, we assume that UniverseMachine provides a complete catalog with galaxy redshifts, SMs, and SFRs that are perfectly known.[10] We draw a million galaxies from a uniform in comoving volume distribution for redshifts less than 1.4 and bin these draws into the redshift bins provided by UniverseMachine. For each galaxy, we draw SM (down to $10^8 M_\odot$) and SFR values from distributions contained in each redshift bin. The mock catalog thus created contains physical distributions (i.e., mass, SFR, redshift distributions) consistent with UniverseMachine, but does not include effects of galaxy clustering. We have also tested the analysis below using the second version of the MICECAT mock galaxy catalog (J. Carretero et al. 2015; M. Crocce et al. 2015; P. Fosalba et al. 2015a, 2015b; K. Hoffmann et al. 2015, 2022), which includes galaxy clustering as well as SM and SFR values, and find similar results to what we report below (see Appendix A). G. Perna et al.(2024) find similar conclusions when investigating potential biases using the first version of the MICECAT catalog using galaxy luminosities.

The distributions for our mock catalog are shown in Figure 1. We see that weighting by SM (green) or SFR (blue) leads to different redshift distributions. As mentioned in A. Vijaykumar et al. (2024), if host galaxies are weighted solely by their SFRs, the redshift evolution of the galaxy number density is roughly proportional to $\sim(1+z)^{2.5}$, whereas solely weighting by their total SMs would be roughly proportional to $\sim(1+z)^{-0.65}$ in the range of redshifts we consider. We also see from the two-dimensional distribution in Figure 1 that SM and SFR are positively correlated with each other. However, due to the two galaxy branches corresponding to star-forming and quiescent galaxies (see, e.g., A. Vijaykumar et al. 2024), although weighting with SFR conserves the positive correlation with SM, weighting with SM leads to a split of hosts between the two branches. We will comment more on this asymmetry in Section 3.

In our simulations, we first randomly assign each galaxy in our mock catalog to a line of sight, with each line of sight containing $N_{\rm GAL}$ galaxies. Doing this effectively gives us a constant number density of galaxies (or, alternatively, a

[9] We use the publicly available UniverseMachine data set from https://halos.as.arizona.edu/UniverseMachine/DR1/SFR/, created using the Bolshoi–Planck (A. A. Klypin et al. 2011) dark matter simulation box with a side length 250 h^{-1} Mpc and 2048^3 particles.

[10] In reality, these quantities would have observational uncertainties, e.g., photometric redshift measurements have a typical fractional uncertainty of a few percent. These uncertainties would need to be accounted for in the likelihood. Mismodeling the shape of the uncertainty regions could also impact the measurement of H_0 (see C. Turski et al. 2023, for a discussion on mismodelling redshift uncertainties).

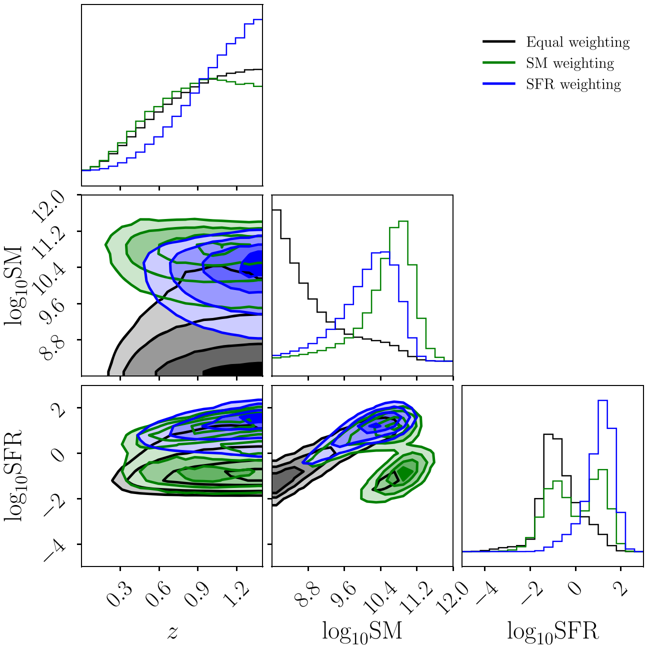

Figure 1. UniverseMachine mock catalog distribution. Notice that the weighted redshift distributions are different—the redshift evolution of the number density follows $\sim(1+z)^{2.5}$ when weighting with SFR and $\sim(1+z)^{-0.65}$ when weighting with SM (A. Vijaykumar et al. 2024). We also note that SFR and SM are correlated, although the correlations change depending on which weighting scheme is used.

constant sky area) along each line of sight. We then generate GW events drawn from our mock catalog weighted by a true (correct) injected weighting scheme. We then calculate an observed luminosity distance for each GW event by scattering the true luminosity distance of the host galaxy by a normal distribution (see Equation (3)). If an observed luminosity distance is positive and less than the threshold luminosity distance, $\hat{d_L}^{\rm thr}$, we say that GW event is detected and take the first $N_{\rm GW}$ detected GW events to use in our inference. We take only the lines of sight that contain GW events and calculate a posterior on H_0 using the method described in Section 2.1, now assuming a different "recovery" weighting scheme. Note that this is slightly different from the procedure in J. R. Gair et al. (2023), where multiple GWs were assumed to all be coming from a single line of sight. This difference more realistically encapsulates that GWs typically come from different lines of sight.[11]

3. Identifying Potential Biases

We follow the prescription outlined in Section 2.1 using three injection sets where (i) all galaxies are equally likely to be hosts, (ii) mergers follow the galaxy SM, and (iii) mergers follow the galaxy SFR. A very conservative uniform H_0 prior of $H_0 = [40, 450]$ km s^{-1} Mpc^{-1} is used. We use an injected H_0 value of $H_0^{\rm inj} = 68$ km s^{-1} Mpc^{-1} to be consistent with the initial parameters of UniverseMachine. Even for the highest value in the prior, the hard edge in our mock catalog at $z = 1.4$ is above the luminosity distance threshold considered. For each injection set, we recover H_0 using the three weighting schemes described above.

An example posterior on H_0 for all injection-recovery schemes is shown in Figure 2. In these results, we draw 20,000 GW events contained within lines of sight, where each line of sight contains 10,000 galaxies. Each GW event has a 20% error (corresponding to $A = 0.2$ in Equation (3)) in luminosity distance. We use 20,000 GW detections to obtain a tight convergence on the final H_0 posterior. Using a smaller subset of detections preserves the magnitude of the bias (albeit with larger statistical uncertainty), but simply adds scatter around the final H_0 posterior obtained using the 20,000 detections. Figure 2 shows that if we assume the correct weighting scheme, we remain unbiased in our estimate of H_0. However, if we assume the incorrect weighting scheme, very large biases become apparent. One striking result is that when the true distribution follows the galaxy SFR, recovering with equal weights results in a bias. We also see that if the true distribution follows SFR, weighting with SM remains unbiased for these initial conditions. However, if the true distribution of mergers follows SM, weighting with SFR is extremely biased. This is due to the asymmetry in the weighted SM–SFR correlations in the UniverseMachine catalog we discussed in Section 2.2. As we see in the two-dimensional SM–SFR distribution in Figure 1, weighting with SFR maintains the positive correlation between SM and SFR in the star-forming branch, which leads to an unbiased estimate of H_0 if we were to incorrectly assume the galaxies follow a SM-weighted distribution due to this positive correlation. On the other hand, if the true host probabilities follow SM, we see in Figure 1 that the most likely galaxy hosts now split between the star-forming and quiescent branches, disrupting the purely positive correlation between SM and SFR, and leading to a bias in H_0 when incorrectly assuming the hosts follow SFR. We discuss the effects behind the biases seen in Figure 2 in more depth in the following subsections, as well as in Appendix B.

We also investigate the influence of uncertainties by looking at cases where GW luminosity distance measurements have a 1%, 10%, or 20% error, as well as the influence of the number of galaxies in our lines of sight, $N_{\rm GAL}$. Figure 3 shows the resulting H_0 posteriors for each injection-recovery set for different runs with different values of $N_{\rm GAL}$, for all fractional errors considered. We discern three different regimes that may lead to significant systematic biases in the inference. We denote these three regimes as the "well-localized" regime occurring at low $N_{\rm GAL}$, the "transitional" regime occurring at moderate $N_{\rm GAL}$, and the "uninformative" regime occurring at large $N_{\rm GAL}$, and describe the results in Figure 3. In Appendix B, we create four other mock catalogs, each removing one feature of UniverseMachine at a time: differing weighted galaxy redshift distributions, correlations between SM and SFR, low amounts of highly weighted galaxies, and the volumetric effects of having a three-dimensional universe. We see how each of these features change the biases in these three regimes to identify which effects determine the biases we see. We describe in more detail these three identified regimes below, but leave further details on how these regimes were identified to Appendix B.

[11] Note that when galaxies also have measurement uncertainties in redshift, J. R. Gair et al. (2023) demonstrate that having a similar number of events to galaxies along a single line of sight leads to a bias. Our toy model does not consider galaxy redshift uncertainties, and thus this potential bias is not relevant to our results. Nonetheless, our prescription will not be impacted even if galaxy redshift uncertainties are considered as we always ensure each line of sight contains many more galaxies than GW events.

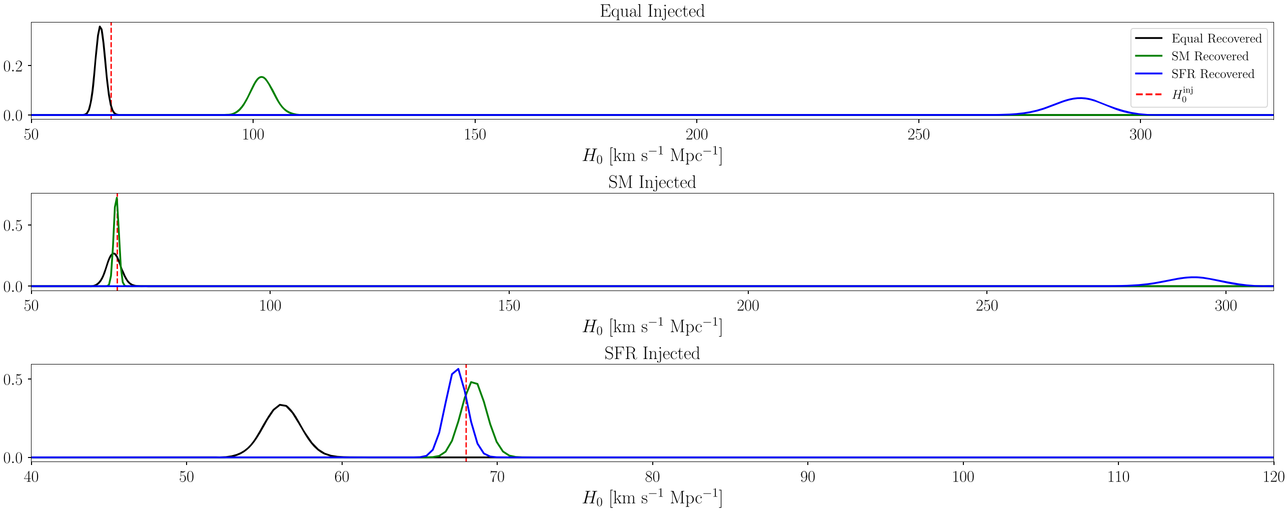

Figure 2. Inference in H_0 using the UniverseMachine mock catalog from 20,000 GW events coming from lines of sight each containing 10,000 galaxies. Each GW event has a 20% error ($A = 0.2$) in luminosity distance. Assuming the correct weighting scheme leads to an unbiased H_0 recovery, but assuming an incorrect weighting scheme can lead to extremely large biases.

3.1. Uninformative Regime

When in a regime where the number of galaxies along each line of sight is high ($N_{\rm GAL} \gtrsim 10^5$; see Figure 3), the posteriors in H_0 fall into the "uninformative" regime. When in this regime, any information from any one galaxy is washed out, and all the information seen in the H_0 posteriors comes from matching the total observed GW luminosity distance distribution with the assumed weighted galaxy redshift distribution (see, e.g., X. Ding et al. 2019; C. Ye & M. Fishbach 2021). To demonstrate this effect, we compute the residual (difference) between the GW observed luminosity distance cumulative distribution function (CDF; $A = 0$) and the weighted galaxy luminosity distance distribution (converted from the redshift distribution for a range of H_0 values) CDF for luminosity distances between zero and $\hat{d}_L^{\rm thr}$. We plot these residuals in Figure 4. When we assume the true (injected) weighting scheme, the GW luminosity distance distribution will follow the injected weighted galaxy redshift distribution, and we recover the correct H_0 value. However, when we weight the galaxy redshifts with the incorrect weighting scheme, the GW luminosity distance distribution will not match up with the weighted redshift distribution for the correct H_0 value (see the red lines in Figure 4), but will match up for some different value of H_0 (see $H_0^{\rm match}$ and the light blue, light green, and gray dashed lines in Figure 4). In Figure 3, for each injected distribution, we plot these best "matched" H_0 values for equal, SM, and SFR recovery weights as gray, light green, and light blue horizontal lines, respectively. As seen in Figure 3, as the number of galaxies along each line of sight increases, the H_0 posteriors for each recovery weighting scheme trend toward the horizontal lines indicating what value of H_0 matches the GW luminosity distance distribution to the weighted galaxy redshift distributions. Likewise, as is shown in Figure 3, for the same value of $N_{\rm GAL}$, increasing the fractional error of the GW luminosity distance measurements also follows the same trend.

3.2. Well-localized Regime

Now we consider the opposite limit—the "well-localized" regime where $N_{\rm GAL}$ is small. Note that in the limit where there is only one galaxy along each line of sight ($N_{\rm GAL} = 1$), we de facto identify the host galaxy and therefore are in effect pursuing the "bright siren" method. Even in the limit where $N_{\rm GAL} > 1$ but below a certain threshold ($N_{\rm GAL} \lesssim 100$), as is apparent in Figure 3, we see that there are no biases regardless of what weighting scheme is used during the inference. The reason for this depends on what the true (correct) weighting scheme is. First, let us consider the case where galaxies all have an equal probability of hosting GW events. In this regime, we find that, on average, there are very few lines of sight containing GW events that contain any extremely highly weighted galaxies (i.e., galaxies with large SM or SFR), such that when combining posteriors, our final inference on H_0 is unbiased for all recovery weighting schemes, regardless of whether it is the correct scheme.

This argument is especially visible in a mock catalog, Uniform:UMuncorrelated, that we introduce in Appendix B. In this catalog, we remove volumetric effects and assign the "stellar mass" to be either 1 or 1000, with 1% of galaxies having a weight of 1000 in order to clearly identify which galaxies are considered "highly weighted" and isolate the bias due only to these highly weighted galaxies. In this case, each line of sight will have $\mathcal{O}(1)$ high-mass galaxy if $N_{\rm GAL} \gtrsim 100$. The "well-localized" regime in this case is when $N_{\rm GAL} \lesssim 100$, where we see in Figure 8 that $N_{\rm GAL} \lesssim 100$ does recover an unbiased H_0 estimate for all injection-recovery combinations for the Uniform:UMuncorrelated mock catalog.

This trend is more difficult to see in more realistic mock catalogs, such as our UniverseMachine catalog, where the galaxy properties have smooth one-dimensional distributions. However, we can provide a rough estimate of the "well-localized" regime limit by the following logic. Let us assume that our mock catalog has ϵ percent of galaxies that have x orders of magnitude higher weights (than the majority of galaxies). Then, we know that there will be, on average, no highly weighted galaxy along each line of sight if $N_{\rm GAL} \lesssim 100/\epsilon$.[12] However, for any bias from a highly weighted galaxy to be relevant, that galaxy would need to have sufficient weight to overwhelm the posterior support from

[12] "Highly weighted" in this case means a galaxy with a weight above 10^x times more than the majority of galaxies.

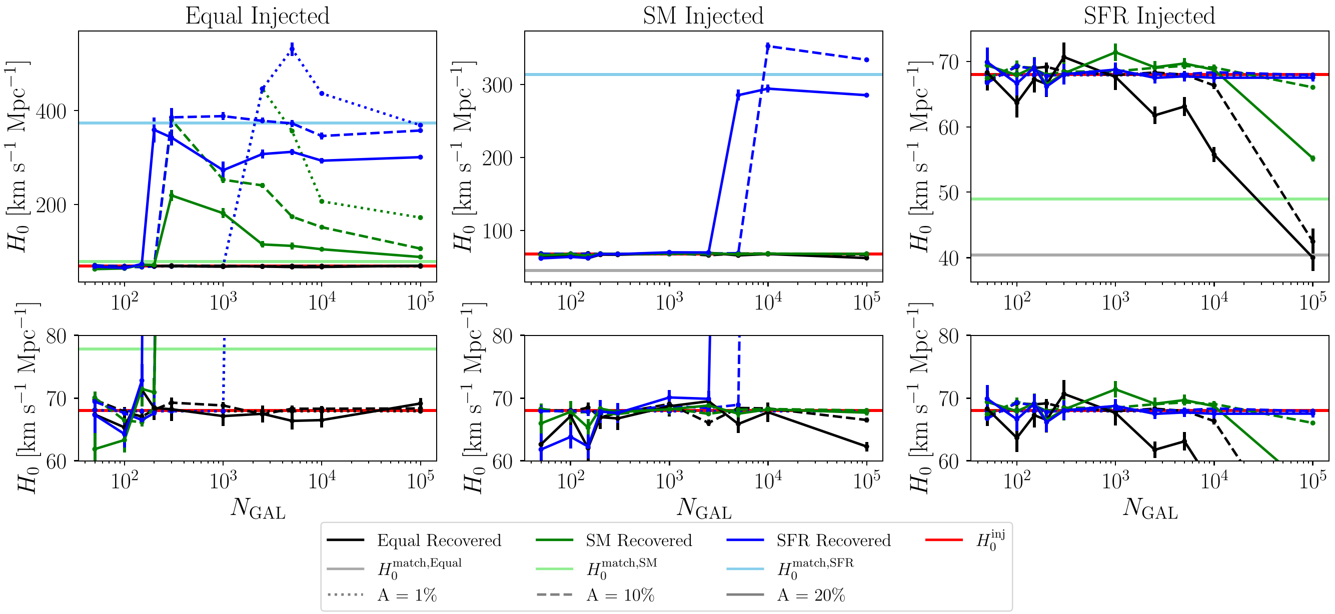

Figure 3. H_0 recovery as a function of $N_{\rm GAL}$ for three choices of fractional error, A, for all injection-recovery weighting schemes. The error bars correspond to the 1σ uncertainty in the posteriors when observing $2 \times N_{\rm GAL}$ GW events. We see that, for small enough $N_{\rm GAL}$, the H_0 inference remains unbiased for all recovery weighting schemes, and for large enough $N_{\rm GAL}$, the biases asymptote to the theoretical best "matched" H_0 values (horizontal lines) from matching up the GW luminosity distance distribution with the weighted galaxy redshift distribution (see the titles in Figure 4). These regimes correspond to the "well-localized" and "uninformative" regimes described in the main text. Decreasing the fractional error shifts the end of the "well-localized" regime and beginning of the "uninformative" regime to higher $N_{\rm GAL}$.

the rest of the galaxies along the line of sight. We use this logic to combine the above condition with the condition that $N_{\rm GAL} \lesssim 10^x$. We can illustrate this with a back-of-envelope calculation using the SM distribution in the UniverseMachine mock catalog. If we assume that the true distribution of GW events is equally likely to be in any host galaxy, the majority of host galaxies will have SM $\sim 10^8 M_\odot$, with $\sim$0.55% of host galaxies having SM $\sim 10^{11} M_\odot$. However, if we weight following SM, the majority of host galaxies will have SM $\sim 10^{11} M_\odot$ (see, e.g., Figure 1). Then, with the above estimate, we would expect the "well-localized" regime to end around $N_{\rm GAL} \sim \frac{100}{0.55} \sim 180$. As we see in the left panel of Figure 3, this estimate is fairly accurate in predicting where the biases begin to appear.

Let us now consider the case where host galaxies follow either SM or SFR weights. When generating GW events, most will have large SM (or SFR) values. Therefore, when using the lines of sight that contain GW events to find a posterior on H_0, each line of sight will have on the order of one large SM (or SFR) galaxy if $N_{\rm GAL} \lesssim 100/\epsilon$ as before, which is typically the true host. When recovering with any weighting scheme, since there are so few galaxies along each line of sight, the true host will always have nonnegligible support. Thus, in this case, the usual argument of dark sirens applies and the true H_0 value will appear from combining many observations.

3.3. Transitional Regime

Finally, in between the "uninformative" and "well-localized" regimes, there is a third regime we have termed the "transitional" regime. In this regime, regardless of the true host-galaxy weighting scheme, any line of sight will have on the order of a few highly weighted galaxies. The $N_{\rm GAL}$ range that determines this regime depends on the percentage of highly weighted galaxies in our mock catalog as well as the fractional error in the GW luminosity distance. In the case of our mock UniverseMachine catalog, this regime corresponds to $N_{\rm GAL} \sim$ 100–10,000 when the GW luminosity distances have a 10% error (see Figure 3). Along a given line of sight, these few extremely large galaxies have weights that are several orders of magnitude higher than other galaxies, and thus they dominate the posterior for that line of sight. If the GW distribution follows SM or SFR, the true host in any line of sight will likely be in one of these few highly weighted galaxies. If we recover with equal weights in this regime, there is still some support for the correct host along each line of sight, and the usual dark siren argument holds such that we will still recover the correct H_0. On the other hand, if the true distribution is such that all galaxies in the catalog are equally likely to be hosts, recovering with SM or SFR leads to extremely large biases. This effect when moving from the "well-localized" to the "transitional" regime can be seen by the very sharp jumps in H_0 bias in Figure 3, especially in the leftmost plot. These potentially large biases appear when we consider that all galaxies along the line of sight contribute to the final H_0 posterior, regardless of whether the galaxy falls within the localization volume of the GW event (or below $\hat{d}_L^{\rm thr}$) for the injected H_0 value. This effect is easiest to see again in the Uniform:UmUncorrelated mock catalog. In this catalog, galaxies are distributed uniformly along a line in redshift. A $\hat{d}_L^{\rm thr}$ of 1550 Mpc corresponds to a redshift of $z \sim 0.3$. However, since our galaxy catalog extends to a redshift of 1.4, when highly weighted galaxies are distributed along the line of sight, the majority of these galaxies will be at a redshift $z > 0.3$. When inferring H_0, these galaxies will give support for $H_0 > H_0^{\rm inj}$, leading to a bias to high H_0 values. While the Uniform:UmUncorrelated mock catalog does not contain any volumetric effects (this catalog only considers a one-dimensional line in distance), when considering more realistic

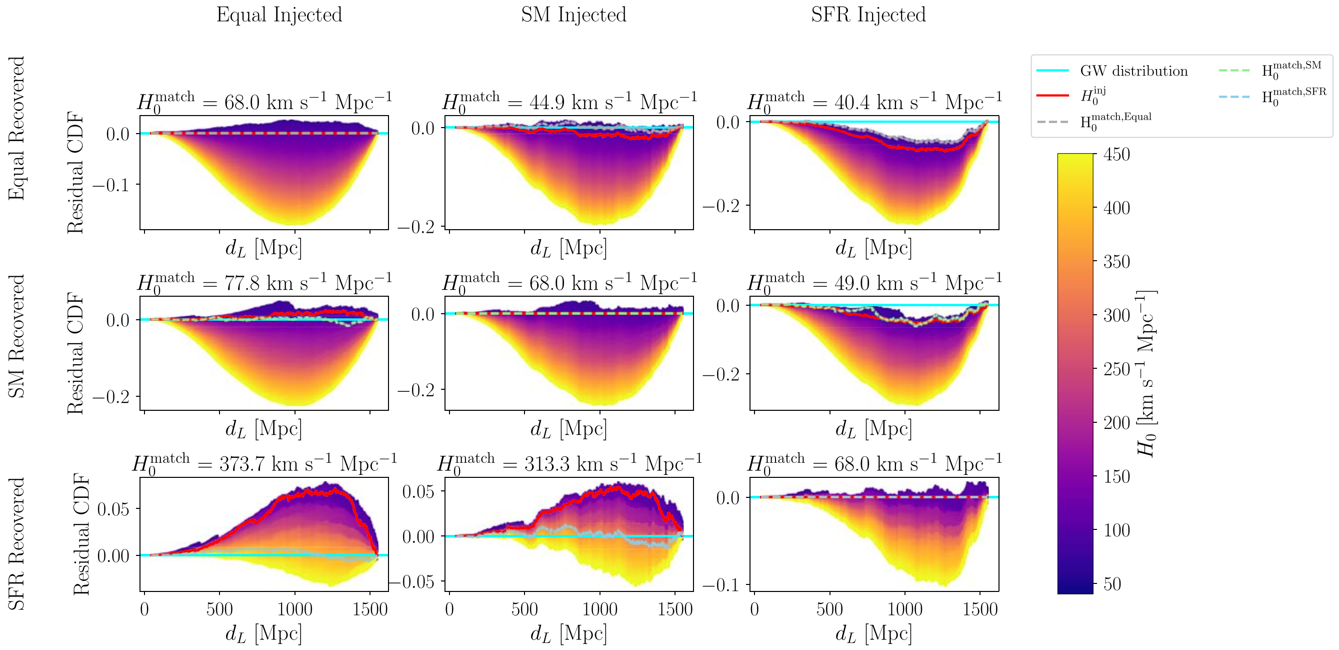

Figure 4. Residual CDFs when comparing a measured GW luminosity distance distribution (with $A = 0$) following a given true injected distribution (cyan; residual CDF = 0) with an assumed weighted galaxy redshift distribution, converted to a luminosity distance distribution for a given value of H_0. The comparison when using the injected value of $H_0^{\rm inj}$ is shown in red for all injections and recoveries. The H_0 value that minimizes the residual, $H_0^{\rm match}$, for each combination is reported in the titles and denoted by the gray, light blue, and light green dashed lines corresponding to a recovery weighting scheme following equal, SFR, or SM weights.

mock catalogs such as our UniverseMachine mock catalog, this argument still holds.

4. Discussion of Potential Biases

In Section 3, we found that assuming an incorrect galaxy host probability distribution can lead to large biases in the recovered value of H_0. In this section, we discuss potential factors that may influence these biases, as well as demonstrate a diagnostic that can be used to identify when an incorrect weighting scheme is used.

4.1. Improving GW Localization

As seen in Figure 3, as we increase the fractional error in GW luminosity distance, we approach the "uninformative" regime. Likewise, this means that decreasing the fractional error will allow us to stay in the "well-localized" regime for longer (i.e., to higher $N_{\rm GAL}$). In other words, if we have better GW localization, there is a better chance that we may be in the "well-localized" regime and recover unbiased estimates of H_0. Intuitively, this is the same as keeping the same fractional error and decreasing the number of galaxies in the localization volume. Therefore, with future GW detectors, e.g., Cosmic Explorer (M. Evans et al. 2021) or the Einstein Telescope (M. Maggiore et al. 2020), we may be able to mitigate possible biases if we only consider events with small GW luminosity distance uncertainties or sky localization such that we are in the "well-localized" regime. However, implementing such a selection would give rise to a nontrivial $P_{\rm det}^{\rm GW}$, making it difficult (but not impossible) to correct for selection biases.

4.2. Exploiting Correlations

As seen in Appendix B, the correlation between SM and SFR helps mitigate the bias when injecting with one and recovering with the other. If there were to be no correlation between SM and SFR, as is the case in the UM:UMuncorrelated catalog in Appendix B, we see that the biases in these two injection-recovery cases are much larger. Since there are correlations in parameter space in actual galaxy catalogs, it is possible that weights built using some optimal combination of galaxy properties could alleviate these biases. For example, if we assume UniverseMachine correctly depicts our Universe, using SM weights may be the best choice to minimize any biases, as long as true host galaxies do not have equal weights. However, a deep knowledge of such a catalog would be needed to build a weighting scheme that exploits these correlations, which is not currently known with present catalogs.

4.3. Higher Signal-to-noise Ratio Cut

In our toy prescription, a lower GW detection threshold $\hat{d}_L^{\rm thr}$ corresponds to a higher signal-to-noise ratio cut. One would expect that, due to GW luminosity distances having a constant fractional error, decreasing the detection threshold would lead to better-localized events. As discussed in Section 4.1, better localization corresponds to staying in the "well-localized" regime longer, but can lead to larger biases in the "transitional" regime. In Figure 5, we investigate the level of biases seen for different injection-recovery schemes when there are 5000 galaxies along each line of sight. In general, we see in Figure 5 that sufficiently decreasing the GW detection threshold does help mitigate some biases, although this may not be the case in general (see, e.g., in the top panel of Figure 5; when we inject equal weights, decreasing the detection threshold from the original 1550 Mpc increases the biases until the threshold reaches about 800 Mpc, where it then starts decreasing).

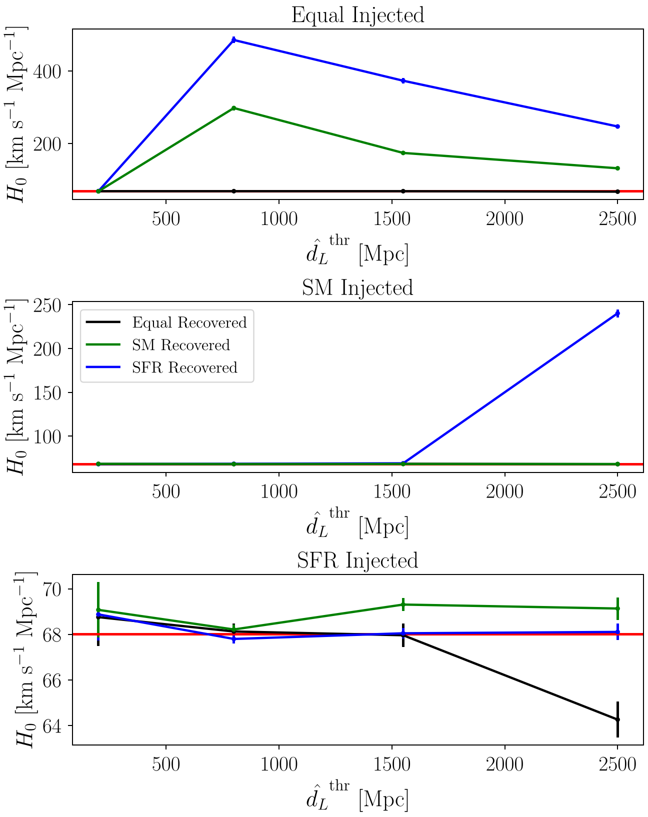

Figure 5. Bias in H_0 inference when changing the GW detection threshold, $\hat{d}_L^{\text{thr}}$. The posteriors were generated from 10,000 GW events coming from lines of sight containing 5000 galaxies. The GW events each have a 10% fractional error in luminosity distance.

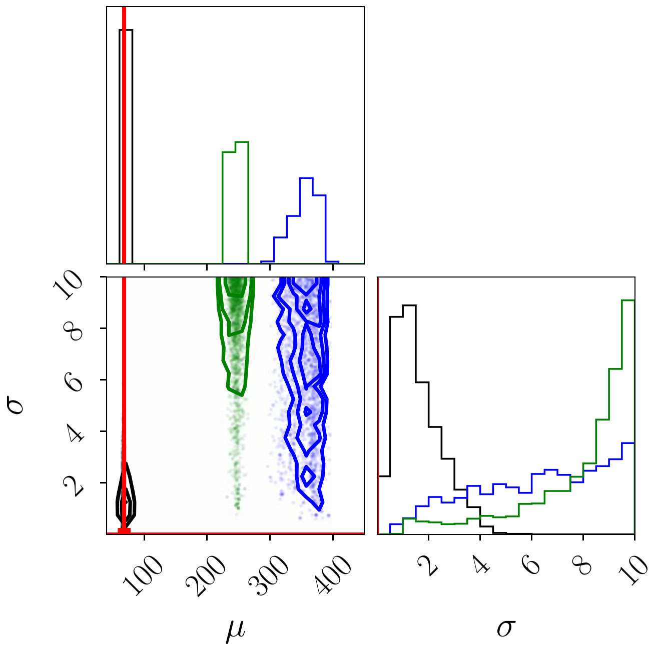

Figure 6. Example posteriors on H_0 inference using hierarchical analysis assuming H_0 is a Gaussian with mean μ and standard deviation σ. The true distribution assumes that host galaxies are all equally likely to be hosts. Each line of sight contains 2500 galaxies and we observe 2500 GW events, each with 10% error in luminosity distance. We see that recovering with SM or SFR causes the recovered standard deviation to shift away from zero, and the mean shifts away from the true value ($H_0^{\text{inj}} = 68$ km s^{-1} Mpc^{-1}). This suggests that using the population of GW events can diagnose when an incorrect weighting scheme is used (and can help "self-calibrate" the correct weighting scheme, thereby mitigating the bias due to incorrect galaxy weighting; see Section 4.5 and A. Vijaykumar et al. 2024).

4.4. Galaxy Clustering

We have not considered the effect of galaxy clustering in this analysis, although it might help mitigate the magnitude of the biases when using an incorrect weighting scheme. While we cannot make any claims on the complete effects of large-scale structure, the mitigating effects of parameter correlations (eg. SM and SFR correlations) seem to, at first order, help decrease the magnitude of biases we see. This supports the claim that further correlations may, again, decrease the biases. However, G. Perna et al. (2024) demonstrate using the MICECAT catalog—a mock catalog that accounts for galaxy clustering—that there are still biases in the inference of H_0 when assuming an incorrect weighting scheme when there are anisotropies in the galaxy structure. J. R. Gair et al. (2023) mention that, over different lines of sight, the over/underdensities in catalogs that contain anisotropic structure would average out. Thus, as seen in G. Perna et al. (2024), there are still biases present due to mismatching the weighted galaxy redshift distributions. We also investigate potential biases using the MICECAT mock catalog using our scheme above in Appendix A and find similar biases to our above results. On the other hand, it is possible that if different weighting schemes prefer the same structure, this positive correlation may lead to an unbiased estimate of H_0, and may even improve it.

4.5. Diagnosing Incorrect Assumptions

We assume a single value of H_0, which allows us to use the methods described throughout this text in our inference. However, in the event that we use an incorrect galaxy host weighting scheme, our analysis need not find that a common value of H_0 is recovered by the population of GW events. As shown in A. Zimmerman et al. (2019), multiplying the individual event likelihoods on H_0 in the latter case fails to capture the deviations present in the sample. Therefore, we suggest using hierarchical analysis (such as in M. Isi et al. 2019) to recover H_0 in order to test for model misspecification such as an incorrect galaxy weighting scheme. For instance, we could posit that instead of the true value of H_0 being a constant, the value of H_0 is drawn from a Gaussian distribution with mean μ and standard deviation σ. If we recover using the correct weighting scheme, we expect to recover a delta function with $\sigma = 0$ and $\mu = H_0^{\text{inj}}$. However, if we recover using the incorrect weighting scheme, we may see that our recovered H_0 value does not converge, or converges with a nonzero standard deviation. An example plot is shown in Figure 6 demonstrating this effect. We inject equal weights for the host-galaxy probabilities, and find that recovering with equal weights indeed results in a posterior at $H_0^{\text{inj}} = 68$ km s^{-1} Mpc^{-1} with support for $\sigma = 0$.[13] However, recovering with weights following either SM or SFR leads to posteriors that have

[13] The posterior peaks slightly away from $\sigma = 0$ while assuming the correct recovery weighting scheme. This is likely because we are trying to probe a very narrow feature (essentially a Dirac delta function) with hierarchical inference while using a finite number of samples to construct the relevant Monte Carlo sums (see, e.g., R. Essick & W. Farr 2022, for a description of this effect). Increasing the number of samples does shift the peak toward $\sigma = 0$, but also increases the computational cost.

nonzero standard deviations and peak away from the injected H_0 value. While this method is extremely useful in diagnosing if an incorrect weighting scheme is used, we emphasize that using hierarchical analysis does not diminish the bias, nor does it give any information on what the correct weighting scheme should be. However, it should be possible to simultaneously infer the weighting scheme as well as H_0 by generalizing the idea laid out in A. Vijaykumar et al. (2024); however, we do not investigate this here.

4.6. *Decreasing the* H*0* *Prior*

We emphasize that our above analysis uses a conservative prior of $H_0 = [40, 450]$ km s^{-1} Mpc^{-1}. The current landscape of Hubble constant measurements, e.g., from measurements of the cosmic microwave background (Planck Collaboration et al. 2020), supernovae (D. Scolnic et al. 2022), or the tip of the red giant branch (W. L. Freedman et al. 2020), reliably place measurements of H_0 to be within ∼60–80 km s^{-1} Mpc^{-1}. While we see that restricting our H_0 prior to these values in the current analysis does not change the biases we see (the posteriors simply rail against the prior bounds), if we use a smaller prior along with the hierarchical inference of H_0 suggested in Section 4.5, the correct H_0 value may be inferred, albeit with some nonzero standard deviation, due to the restrictive prior absorbing some of the systematic uncertainty present when using an incorrect weighting scheme. While using hierarchical analysis with a very restrictive H_0 prior may yield unbiased results, it may fail if individual GW event posteriors are largely uninformative over this smaller prior range. In that case, the hierarchical analysis would infer a mean at the center of the prior range and a large standard deviation that encompasses the entire prior. Therefore, while we recommend always using hierarchical analysis in future inferences due to its strength in diagnosing incorrect weighting schemes, incorporating a restrictive H_0 prior may additionally yield unbiased results regardless of weighting scheme. However, we emphasize that additional investigations need to be carried out to confirm if this is always true.

5. Conclusion

We have examined the dark siren approach to cosmology, wherein all galaxies in a binary's localization volume are considered as potential hosts to a given source. In particular, we have explored the impacts of weighting the galaxy catalog incorrectly, and find the potential for substantial biases in the inferred value of H_0. We break these biases into three regimes determined by the number of galaxies in each line of sight, N_{GAL}: the "well-localized," "transitional," and "uninformative" regimes. We create multiple galaxy catalog toy models to isolate each effect that might influence the observed biases. We advocate the use of hierarchical analysis during H_0 inference to help diagnose any potential biases, as any noticeable standard deviation in the H_0 posterior identifies the use of an incorrect weighting scheme. We find that correlations between parameters such as SM and SFR, as well as correlations with large-scale structure, may reduce potential biases. We also note that future GW detectors that improve GW luminosity distance localization may help mitigate some of these biases, so long as the number of galaxies along any given line of sight remains small. Note that our results also assume a 100% complete galaxy survey; a realistic survey will be incomplete, and the choice of weights should also be taken into account while correcting for catalog incompleteness.

Finally, we note that current LVK analyses (R. Abbott et al. 2023b) do not find any substantial bias in the recovered H_0 value compared to conventional, non-standard siren determinations. There are a number of reasons for this unbiased determination. First, as reported by R. Abbott et al. (2023b), current constraints of H_0 from galaxy catalogs with K-band weighting give a ∼18% measurement, while a spectral siren analysis using a fixed GW population without galaxy catalogs yields a ∼20% measurement of H_0. This demonstrates that LVK analyses using the dark siren method are presently dominated by uncertainties associated with the GW population, and hence any incorrect weighting scheme when using galaxy catalogs would be expected to have a subdominant effect. In addition, the current luminosity weighting schemes may be well informed (see, e.g., A. Vijaykumar et al. 2024, which constrains host galaxies from the evolution of the GW merger rates), such that current LVK analyses may be using a sufficiently accurate weighting so that the bias is minimized. Additionally, a joint inference of the binary population properties and H_0 (R. Gray et al. 2023; S. Mastrogiovanni et al. 2023) could help mitigate this bias, although consistency should be ensured between the galaxy weights and the redshift evolution of the merger rate (A. Vijaykumar et al. 2024) while constructing the likelihood function. Finally, large-scale clustering in the galaxy catalogs currently in use by the LVK may also help to mitigate these biases. Even so, as our data are improved, the bias in H_0 due to incorrect galaxy weighting may become an increasing concern for dark siren approaches. Our work highlights the importance of accounting for and mitigating the bias due to incorrect galaxy weighting in future dark siren measurements.

Acknowledgments

We thank the anonymous referee for a very careful review of our manuscript. We also thank Gabriele Perna for helpful suggestions. A.G.H. is supported by NSF grants AST-2006645 and PHY-2110507, and gratefully acknowledges the ARCS Foundation Scholar Award through the ARCS Foundation, Illinois Chapter, with support from the Brinson Foundation. A.V. and M.F. acknowledge support to CITA by the Natural Sciences and Engineering Research Council of Canada (NSERC) (funding reference number 568580). A.V. is also supported by a Fulbright Program grant under the Fulbright-Nehru Doctoral Research Fellowship, sponsored by the Bureau of Educational and Cultural Affairs of the United States Department of State and administered by the Institute of International Education and the United States-India Educational Foundation. D.E.H. is supported by NSF grants 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. This work has made use of CosmoHub (J. Carretero et al. 2017; P. Tallada et al. 2020).

Software: numpy (C. R. Harris et al. 2020), scipy (P. Virtanen et al. 2020), matplotlib (J. D. Hunter 2007), astropy (Astropy Collaboration et al. 2013, 2018; Astropy Collaboration 2022), jupyter (T. Kluyver et al. 2016), pandas (W. McKinney 2010), jax (J. Bradbury et al. 2018), numpyro (D. Phan et al. 2019; E. Bingham et al. 2019).

Appendix A
Example of Bias Using the MICECAT Mock Galaxy Catalog

To investigate the effects of galaxy clustering on potential biases in H_0 inference if the incorrect host-galaxy weighting scheme is used, we use the same prescription as outlined in Section 2.2, now using the second version of the MICECAT mock galaxy catalog. Figure 7 demonstrates the potential bias in H_0 when using the MICECAT catalog. In this analysis, each line of sight contains 10,000 galaxies, with 20,000 GW events observed, each with a fractional error of 20%. These are the same parameters as used for the UniverseMachine analysis in Figure 2. When comparing with Figure 2, we see that using the MICECAT catalog yields similar biases as when using the UniverseMachine mock catalog. One difference is that, due to the greater discrepancy in the MICECAT catalog between the equally-weighted galaxy redshift distribution with the SM-weighted and SFR-weighted redshift distributions, injecting with either SM or SFR weights and recovering with equal weights leads to a larger bias than is seen in Figure 2. Another difference is that, while the magnitude of the bias in the SM injected–SFR recovered case is slightly less severe, the SFR injected–SM recovered bias becomes more severe. We note that this is only a proof of concept to demonstrate that biases using the MICECAT catalog are similar to those in the UniverseMachine catalog and not a full analysis of how galaxy clustering affects the results presented above.

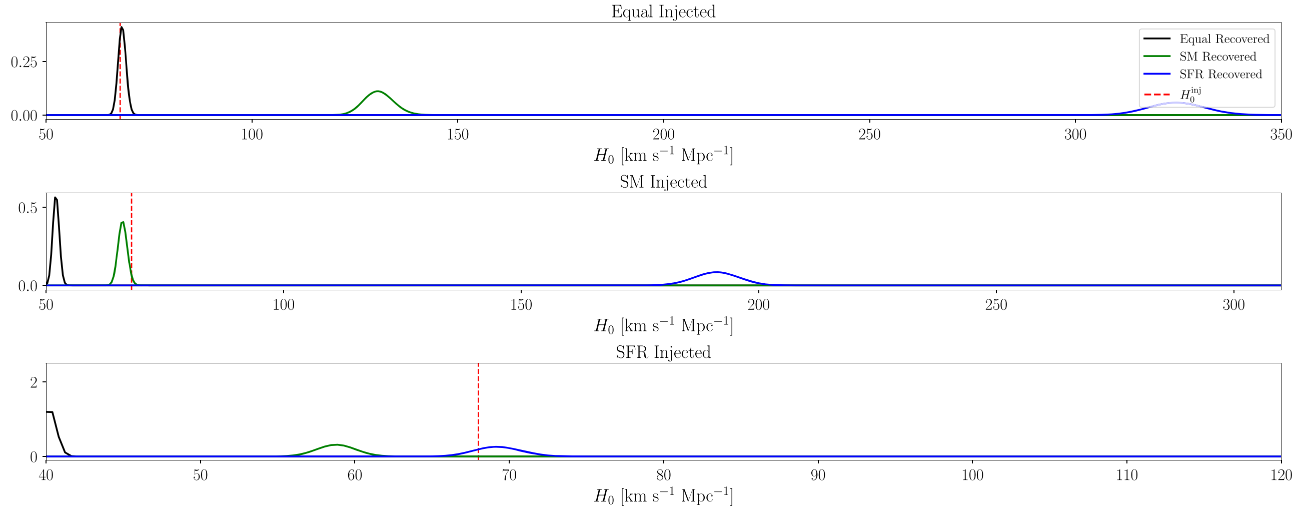

Figure 7. Inference in H_0 using the MICECAT mock catalog from 20,000 GW events, with each line of sight containing 10,000 galaxies. Each GW event has a 20% error ($A = 0.2$) in luminosity distance. Assuming the correct weighting scheme leads to an unbiased H_0 recovery, but assuming an incorrect weighting scheme can lead to extremely large biases. The magnitudes of the biases seen here are similar to those seen when using the UniverseMachine mock catalog (see Figure 2).

Appendix B
Discovering Trends in the Systematic Biases in H_0 Inference Using Simple Mock Catalogs

To isolate different effects on the systematic biases seen in Section 3, we create four more mock catalogs, which decrease in levels of complexity. The catalogs are created by drawing a million galaxies from various redshift distributions for redshifts less than 1.4. However, the four catalogs differ in a couple of ways, and are defined as follows:

1. UM:UMuncorrelated: This catalog is created in the same way as is described in Section 2.2, but now SM and SFR are sampled from their marginalized one-dimensional distributions given by UniverseMachine such that there is no longer a correlation between SM and SFR but each have the same weighted redshift distributions as in the original catalog.
2. CC:UMcorrelated: Galaxies are drawn from a constant in comoving volume distribution, while SMs and SFRs are assigned to each galaxy following the two-dimensional SM and SFR distribution given by UniverseMachine, such that SM and SFR are correlated with each other but not correlated in redshift.
3. CC:Uniform: Galaxies are drawn from a constant in comoving volume distribution, and SMs and SFRs are assigned independently to each galaxy, both following uniform distributions between the minimum and maximum value given by the UniverseMachine SM and SFR distributions.
4. Uniform:UMuncorrelated: This mock catalog ignores volumetric effects by distributing galaxies uniform in redshift in a one-dimensional Euclidean universe. Galaxies are assigned an SFR drawn from the marginalized one-dimensional SFR distribution given by UniverseMachine, while SM is assigned a weight of 1 or 1000 depending on a given percentage of highly weighted galaxies, $\epsilon = 1\%$.

These mock catalogs are specifically chosen to investigate potentially significant effects independently from the full UniverseMachine mock catalog, such as influences from the correlation between SM and SFR (UM:UMuncorrelated), correlations of SM and SFR with redshift leading to different weighted redshift distributions (CC:UMcorrelated), and low-number statistics of highly weighted galaxies (CC:Uniform). We also investigate any potential volumetric effects by creating a mock catalog that is set in a one-dimensional Euclidean universe (Uniform:UMuncorrelated).

Figure 8 shows posteriors on H_0 for all injection-recovery weighting schemes for three representative total number of galaxies along each line of sight, $N_{\mathrm{GAL}} = [80, 5000, 400, 000]$, indicating the "well-localized," "transitional", and "uninformative" regimes, respectively. The first aspect to note is that, even in the case of a one-dimensional Euclidean universe, recovering with the incorrect weighting scheme using the Uniform:UMuncorrelated catalog yields very large biases in the "transitional" regime but not in the "well-localized" and "uninformative" regimes. In the "well-localized" regime when the true hosts are all equally weighted, on average most of the lines of sight will not have a highly weighted galaxy when recovering with SM or SFR due to the small ϵ percentage of large galaxies. In the "uninformative" regime, the weighted galaxy redshift distributions are all equivalent, and thus we do not expect any bias in this case either. However, in the "transitional" regime, since the galaxies are distributed uniformly in redshift, with a $\hat{d}_L^{\,\mathrm{thr}} = 1550$ Mpc corresponding to a $z \sim 0.3$, most of the highly weighted galaxies will fall above this threshold for the injected H_0 value but will be within the GW localization volume for a large H_0, leading to a bias to high H_0, as we see in Figure 8. We do see that, for this catalog, recovering with equal weights when the true distribution

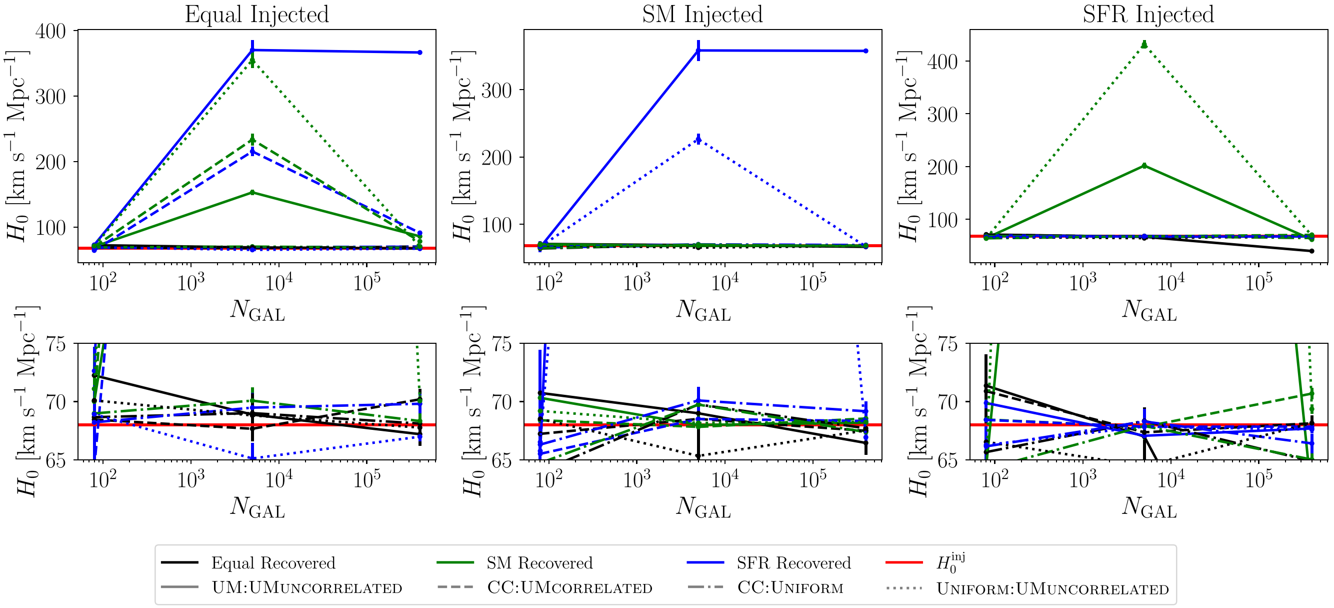

Figure 8. Inference on H_0 for four toy mock galaxy catalogs, UM:UMuncorrelated (solid), CC:UMcorrelated (dashed), CC:Uniform (dashdot), and Uniform:UMuncorrelated (dotted), for three N_{GAL}, representing the "well-localized," "transitional," and "uninformative" regimes. Recovering with equal weights is given by the black lines, recovering with SM is given by the green lines, and recovering with SFR is given by the blue lines. All GW events have a 10% fractional error in luminosity distance. The error bars correspond to the 1σ uncertainty in the posteriors when observing $N_{\mathrm{GAL}}/4$ GW events.

follows SM or SFR in the “transitional” regime remains unbiased. This is because the redshift distributions are the same, and since all the weights are equal for all galaxies along the line of sight, recovering with equal weights will still have support for the correct host galaxy, albeit with less probability due to considering all galaxies along the line of sight.

On the other hand, we also see that all recovery weighting schemes using the CC:UNIFORM remain unbiased regardless of the number of galaxies in each line of sight. This is due to this catalog’s SM and SFR distributions having a very large number of highly weighted galaxies such that we are never in the “transitional” regime. Since the weighted galaxy redshift distributions are all equivalent, we again see no biases in the “uninformative” regime. However, when we consider more realistic SM and SFR distributions, as in the CC:UMCORRELATED catalog, the biases in the “transitional” regime reemerge. In this catalog, SM and SFR are also correlated with each other, and we see that in this case, injecting with one and recovering with the other does not lead to any substantial biases.

The final catalog we consider is the UM:UMUNCORRELATED catalog, which contains the same weighted redshift distributions as the original UNIVERSEMACHINE mock catalog we consider in the main text, but now SM and SFR are no longer correlated with each other. As we see in Figure 8, the biases seen in all regimes match those seen using the UNIVERSEMACHINE mock catalog, but new biases emerge when injecting with either SM or SFR and recovering with the other. This effect demonstrates that correlations between SM and SFR can help mitigate the biases seen in the “transitional” regime.

ORCID iDs

Alexandra G. Hanselman https://orcid.org/0000-0002-8304-0109
Aditya Vijaykumar https://orcid.org/0000-0002-4103-0666
Maya Fishbach https://orcid.org/0000-0002-1980-5293
Daniel E. Holz https://orcid.org/0000-0002-0175-5064

References

Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2017a, PhRvL, 119, 161101
Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2017b, ApJL, 848, L13
Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2017c, ApJL, 848, L12
Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2017d, Natur, 551, 85
Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2021, ApJ, 909, 218
Abbott, R., Abbott, T. D., Acernese, F., et al. 2023a, PhRvX, 13, 041039
Abbott, R., Abe, H., Acernese, F., et al. 2023b, ApJ, 949, 76
Adhikari, S., Fishbach, M., Holz, D. E., Wechsler, R. H., & Fang, Z. 2020, ApJ, 905, 21
Artale, M. C., Mapelli, M., Giacobbo, N., et al. 2019, MNRAS, 487, 1675
Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, ApJ, 935, 167
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33
Behroozi, P., Wechsler, R. H., Hearin, A. P., & Conroy, C. 2019, MNRAS, 488, 3143
Bell, E. F., McIntosh, D. H., Katz, N., & Weinberg, M. D. 2003, ApJS, 149, 289
Bera, S., Rana, D., More, S., & Bose, S. 2020, ApJ, 902, 79
Bingham, E., Chen, J. P., Jankowiak, M., et al. 2019, JMLR, 20, 1
Bradbury, J., Frostig, R., Hawkins, P., et al., 2018 JAX: Composable Transformations of Python+NumPy Programs, 0.3.13, github.com/jax-ml/jax
Carretero, J., Castander, F. J., Gaztañaga, E., Crocce, M., & Fosalba, P. 2015, MNRAS, 447, 646
Carretero, J., Tallada, P., Casals, J., et al. 2017, in Proc. European Physical Society Conf. on High Energy Physics, 314 (Trieste: SISSA), 488
Chatterjee, D., Hegade, K. R. A., Holder, G., et al. 2021, PhRvD, 104, 083528
Chen, H.-Y., Fishbach, M., & Holz, D. E. 2018, Natur, 562, 545
Cigarrán Díaz, C., & Mukherjee, S. 2022, MNRAS, 511, 2782
Coulter, D. A., Foley, R. J., Kilpatrick, C. D., et al. 2017, Sci, 358, 1556
Crocce, M., Castander, F. J., Gaztañaga, E., Fosalba, P., & Carretero, J. 2015, MNRAS, 453, 1513
Cutler, C., & Flanagan, É. E. 1994, PhRvD, 49, 2658
Dalal, N., Holz, D. E., Hughes, S. A., & Jain, B. 2006, PhRvD, 74, 063006
Del Pozzo, W. 2012, PhRvD, 86, 043011
Di Valentino, E., Mena, O., Pan, S., et al. 2021, CQGra, 38, 153001
Ding, X., Biesiada, M., Zheng, X., et al. 2019, JCAP, 2019, 033
Essick, R., & Farr, W. 2022, arXiv:2204.00461
Evans, M., Adhikari, R. X., Afle, C., et al. 2021, arXiv:2109.09882
Ezquiaga, J. M., & Holz, D. E. 2022, PhRvL, 129, 061102
Farr, W. M., Fishbach, M., Ye, J., & Holz, D. E. 2019, ApJL, 883, L42
Finke, A., Foffa, S., Iacovelli, F., Maggiore, M., & Mancarella, M. 2021, JCAP, 2021, 026
Finn, L. S., & Chernoff, D. F. 1993, PhRvD, 47, 2198
Fishbach, M., Gray, R., Magaña Hernandez, I., et al. 2019, ApJL, 871, L13
Fosalba, P., Crocce, M., Gaztañaga, E., & Castander, F. J. 2015a, MNRAS, 448, 2987
Fosalba, P., Gaztañaga, E., Castander, F. J., & Crocce, M. 2015b, MNRAS, 447, 1319
Freedman, W. L., & Madore, B. F. 2023, JCAP, 2023, 050
Freedman, W. L., Madore, B. F., Hoyt, T., et al. 2020, ApJ, 891, 57
Gair, J. R., Ghosh, A., Gray, R., et al. 2023, AJ, 166, 22
Gray, R., Beirnaert, F., Karathanasis, C., et al. 2023, JCAP, 2023, 023
Gray, R., Hernandez, I. M., Qi, H., et al. 2020, PhRvD, 101, 122001
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Hoffmann, K., Bel, J., Gaztañaga, E., et al. 2015, MNRAS, 447, 1724
Hoffmann, K., Secco, L. F., Blazek, J., et al. 2022, PhRvD, 106, 123510
Holz, D. E., & Hughes, S. A. 2005, ApJ, 629, 15
Hunter, J. D. 2007, CSE, 9, 90
Isi, M., Chatziioannou, K., & Farr, W. M. 2019, PhRvL, 123, 121101
Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, Positioning and Power in Academic Publishing: Players, Agents and Agendas (Amsterdam: IOS Press), 87
Klypin, A. A., Trujillo-Gomez, S., & Primack, J. 2011, ApJ, 740, 102
Lamberts, A., Garrison-Kimmel, S., Clausen, D. R., & Hopkins, P. F. 2016, MNRAS, 463, L31
MacLeod, C. L., & Hogan, C. J. 2008, PhRvD, 77, 043512
Maggiore, M., Van Den Broeck, C., Bartolo, N., et al. 2020, JCAP, 2020, 050
Mancarella, M., Finke, A., Foffa, S., et al. 2022, arXiv:2203.09238
Mapelli, M., Giacobbo, N., Santoliquido, F., & Artale, M. C. 2019, MNRAS, 487, 2
Mastrogiovanni, S., Laghi, D., Gray, R., et al. 2023, PhRvD, 108, 042002
Mastrogiovanni, S., Leyde, K., Karathanasis, C., et al. 2021, PhRvD, 104, 062009
McKinney, W. 2010, in Proc. 9th Python in Science Conf., ed. S. van der Walt & J. Millman (SciPy), 56
Messenger, C., & Read, J. 2012, PhRvL, 108, 091101
Mukherjee, S., Wandelt, B. D., Nissanke, S. M., & Silvestri, A. 2021, PhRvD, 103, 043520
Namikawa, T., Nishizawa, A., & Taruya, A. 2016, PhRvL, 116, 121302
Nissanke, S., Holz, D. E., Dalal, N., et al. 2013, arXiv:1307.2638
Nissanke, S., Holz, D. E., Hughes, S. A., Dalal, N., & Sievers, J. L. 2010, ApJ, 725, 496
O’Shaughnessy, R., Kalogera, V., & Belczynski, K. 2010, ApJ, 716, 615
Palmese, A., Bom, C. R., Mucesh, S., & Hartley, W. G. 2023, ApJ, 943, 56
Palmese, A., deVicente, J., Pereira, M. E. S., et al. 2020, ApJL, 900, L33
Perna, G., Mastrogiovanni, S., & Ricciardone, A. 2024, arXiv:2405.07904
Phan, D., Pradhan, N., & Jankowiak, M. 2019, arXiv:1912.11554
Pierra, G., Mastrogiovanni, S., Perriès, S., & Mapelli, M. 2024, PhRvD, 109, 083504
Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, A&A, 641, A6
Rauf, L., Howlett, C., Davis, T. M., & Lagos, C. D. P. 2023, MNRAS, 523, 5719
Santoliquido, F., Mapelli, M., Artale, M. C., & Boco, L. 2022, MNRAS, 516, 3297
Schutz, B. F. 1986, Natur, 323, 310
Scolnic, D., Brout, D., Carr, A., et al. 2022, ApJ, 938, 113
Singer, L. P., Chen, H.-Y., Holz, D. E., et al. 2016, ApJL, 829, L15
Soares-Santos, M., Holz, D. E., Annis, J., et al. 2017, ApJL, 848, L16
Soares-Santos, M., Palmese, A., Hartley, W., et al. 2019, ApJL, 876, L7

Srinivasan, R., Lamberts, A., Bizouard, M. A., Bruel, T., & Mastrogiovanni, S. 2023, MNRAS, 524, 60

Tallada, P., Carretero, J., Casals, J., et al. 2020, A&C, 32, 100391

Taylor, S. R., Gair, J. R., & Mandel, I. 2012, PhRvD, 85, 023535

Toffano, M., Mapelli, M., Giacobbo, N., Artale, M. C., & Ghirlanda, G. 2019, MNRAS, 489, 4622

Trott, E., & Huterer, D. 2023, PDU, 40, 101208

Turski, C., Bilicki, M., Dálya, G., Gray, R., & Ghosh, A. 2023, MNRAS, 526, 6224

Uddin, S. A., Burns, C. R., Phillips, M. M., et al. 2024, ApJ, 970, 72

Vijaykumar, A., Fishbach, M., Adhikari, S., & Holz, D. E. 2024, ApJ, 972, 157

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, NatMe, 17, 261

Ye, C., & Fishbach, M. 2021, PhRvD, 104, 043507

Zimmerman, A., Haster, C.-J., & Chatziioannou, K. 2019, PhRvD, 99, 124044