



Blinded Mock Data Challenge for Gravitational-wave Cosmology. I. Assessing the Robustness of Methods Using Binary Black Hole Mass Spectrum

Aman Agarwal^{1,2,3,18} , Ulyana Dupletsa^{4,5,6,18} , Konstantin Leyde^{7,18} , Suvodip Mukherjee^{8,19} , Benot Revenu^{9,10,20} ,
 Juan Esteban Rivera^{11,20} , Antonio Enea Romano^{11,20} , Mohit Raj Sah^{8,20} , Sergio Vallejo-Peña^{11,20} ,
 Adrian Avendano^{11,21} , Freija Beirnaert^{12,21} , Gergely Dalya^{13,14,21} , Miguel Cifuentes Espitia^{11,22} ,
 Christos Karathanasis^{15,21} , Santiago Moreno- Gonzalez^{11,22} , Lucas Quiceno^{11,22} , Federico Stachurski^{16,21} ,
 Juan Garcia-Bellido¹⁷ , Rachel Gray¹⁶ , Nicola Tamanini¹³ , and Cezary Turski¹²

¹ Perimeter Institute for Theoretical Physics, Waterloo, Ontario, N2L 2Y5, Canada

² Department of Physics, University of Guelph, Guelph, Ontario, N1G 2W1, Canada

³ Institute of Physics, University of Greifswald, 17489 Greifswald, Germany

⁴ Gran Sasso Science Institute (GSSI), I-67100 L'Aquila, Italy

⁵ INFN, Laboratori Nazionali del Gran Sasso, I-67100 Assergi, Italy

⁶ Institute of High Energy Physics—Austrian Academy of Sciences, 1010 Vienna, Austria

⁷ ICG Portsmouth, University of Portsmouth, Portsmouth, Hampshire, UK

⁸ Department of Astronomy and Astrophysics, Tata Institute of Fundamental Research, Mumbai 400005, India; suvodip@tifr.res.in

⁹ Subatech, CNRS—Institut Mines-Telecom Atlantique, Université de Nantes, France

¹⁰ Université Paris-Saclay, CNRS/IN2P3, IJCLab, 91405 Orsay, France

¹¹ Instituto de Física, Universidad de Antioquia, A.A.1226, Medellín, Colombia

¹² Universiteit Gent, Proeftuinstraat 86, 9000 Ghent, Belgium

¹³ Laboratoire des 2 Infinis—Toulouse (L2IT-IN2P3), Université de Toulouse, CNRS, UPS, F-31062 Toulouse Cedex 9, France

¹⁴ MTA-ELTE Astrophysics Research Group, 1117 Budapest, Hungary

¹⁵ Institut de Física d'Altes Energies (IFAE), Barcelona Institute of Science and Technology, Barcelona, Spain

¹⁶ SUPA, University of Glasgow, Glasgow, G12 8QQ, UK

¹⁷ Instituto de Física Teórica IFT-UAM/CSIC, Universidad Autónoma de Madrid, 28049 Madrid, Spain

Received 2025 January 14; revised 2025 May 15; accepted 2025 May 17; published 2025 June 26

Abstract

Gravitational wave sources are standard sirens that provide an independent way to map the cosmic expansion history by combining with an independent redshift measurement, either from an electromagnetic counterpart for a bright siren or using different statistical techniques for dark sirens. In this analysis, we perform the first Blinded Mock Data Challenge (Blinded-MDC) to test the robustness in inferring the value of Hubble constant H_0 for a dark siren technique which depends on astrophysical mass distribution of binary black holes (BBHs). We have considered different analysis setups for the Blinded-MDC to test both statistical and systematic uncertainties and demonstrate the capabilities in inferring H_0 with detector sensitivity as per the fourth observation run of LIGO-Virgo-KAGRA. We find that when the astrophysical population of BBHs matches with the underlying assumption of the model, a cosmological pipeline can recover the injected parameters using the observed mass distribution. However, when the mock mass distribution of the astrophysical population depends slightly on redshift and one is ignorant about it in analyzing the data, it can cause a systematic discrepancy in the inferred value of H_0 by about 1.5σ , above the statistical fluctuations due to noise and a limited number of events. In the future, elaborate studies will be required to mitigate systematic uncertainties due to unknown astrophysical complexities. This MDC framework sets the road map for inspecting the precision and accuracy of standard siren cosmology and provides the first insight into the robustness of the population-dependent cosmology inference in a blinded analysis setup.

Unified Astronomy Thesaurus concepts: Gravitational waves (678); Cosmology (343); Hubble constant (758)

1. Introduction

The detection of gravitational waves (GWs) by the LIGO-Virgo-KAGRA (LVK) Scientific Collaboration has opened a window to the cosmos, making it possible to explore physics

from the smallest scale associated with neutron stars to cosmological scales of the Universe and answering several fundamental questions about the Universe (B. P. Abbott et al. 2016). One such fundamental question is the expansion history of the Universe as a function of cosmological redshift. Such a measurement will not only shed light on the mismatch in the value of the current expansion rate (called the Hubble constant H_0) (L. Verde et al. 2019; E. Abdalla et al. 2022) but will also discover the constituents that contribute significantly to the energy budget of the Universe, such as dark matter and dark energy.

Measurement of the expansion history of the Universe using GW sources is feasible using sources having electromagnetic (EM) counterparts, such as binary neutron stars and neutron star-black holes, and sources without EM counterparts, such as stellar

¹⁸ Editorial Team and Analysis Group.

¹⁹ Coordinator, Editorial Team, and Simulation Group.

²⁰ Editorial Team and Simulation Group.

²¹ Simulation Group.

²² Analysis Group.



origin binary black holes (BBHs) without the presence of any baryonic matter around them. However, the population of BBHs formed in an active galactic nucleus disk can have EM counterparts (H.-Y. Chen et al. 2022; S. Mukherjee et al. 2020b; V. Gayathri et al. 2021; C. R. Bom & A. Palmese 2024). For the former, measuring the redshift of the host galaxy of the GW sources is possible through spectroscopic follow-ups of the EM counterpart, as shown in the case of GW170817 (B. P. Abbott et al. 2017a, 2017b). For the latter, various techniques have been developed and applied to the second (GWTC-2) or third (GWTC-3) GW catalog of the LVK Scientific Collaboration (B. P. Abbott et al. 2021; A. Finke et al. 2021; R. Abbott et al. 2023b; C. Karathanasis et al. 2023a; S. Mukherjee et al. 2024). These techniques include a mass spectrum of BBHs through the spectral-sirens technique (W. M. Farr et al. 2019; S. Mastrogianni et al. 2021; S. Mukherjee 2022; J. M. Ezquiaga & D. E. Holz 2022; K. Leyde et al. 2022; A. M. Farah et al. 2025; K. Leyde et al. 2024; G. Pierra et al. 2024; I. Magaña Hernandez & A. Ray 2024; U. Mali & R. Essick 2025; G. Perna et al. 2025), a statistical host identification technique using galaxy catalogs (B. F. Schutz 1986; M. Fishbach et al. 2019; M. Soares-Santos et al. 2019; R. Gray et al. 2020; A. Palmese et al. 2021; R. Gray et al. 2023), and the GW-galaxy cross-correlation technique for LVK (S. Mukherjee & B. D. Wandelt 2018; S. Bera et al. 2020; S. Mukherjee et al. 2020a, 2021b; J. Ferri et al. 2025), which will be useful in synergy with spectroscopic galaxy surveys such as DESI (DESI Collaboration et al. 2016), Euclid (A. Refregier et al. 2010), and SPHEREx (O. Dore et al. 2018), and the upcoming photometric survey Vera Rubin Observatory (LSST Science Collaboration et al. 2009) as demonstrated in (C. C. Diaz & S. Mukherjee 2022; S. Afroz & S. Mukherjee 2024). Using these techniques, the value of the Hubble constant after combining both bright and dark standard sirens inferred using spectral sirens assuming no redshift evolution of mass distribution is $H_0 = 68^{+12}_{-8} \text{ km s}^{-1} \text{ Mpc}^{-1}$ (R. Abbott et al. 2023b), using statistical host identification assuming a fixed BBH population model is $H_0 = 68^{+8}_{-6} \text{ km s}^{-1} \text{ Mpc}^{-1}$ (R. Abbott et al. 2023b), and using the cross-correlation technique (after marginalizing over BBH bias parameters controlling the population assumption) is $H_0 = 75.4^{+11}_{-6} \text{ km s}^{-1} \text{ Mpc}^{-1}$ (S. Mukherjee et al. 2024).

However, one of the key requirements for an accurate and precise cosmological probe is to understand the impact of astrophysical uncertainty due to various assumptions and the reliability of a technique in providing accurate cosmological results despite these challenges. Such analyses are performed for other cosmological probes to the expansion history, such as the cosmic microwave background (Planck Collaboration et al. 2016), baryon acoustic oscillation (A. Pérez-Fernández et al. 2025), and lensing time delay measurements (X. Ding et al. 2018) to understand the impact from possible known sources of systematics. In this work, we explore for the first time a Blinded Mock Data Challenge (Blinded-MDC) on realistically simulated GW sources to understand the interplay between cosmological inference of H_0 and astrophysical population assumptions of BBHs. This Blinded-MDC analysis is focused on the method that uses the mass distribution of the BBHs to infer the expansion history of the Universe and the impact of the underlying assumption on the spectral-sirens technique. In the future, the Blinded-MDC technique will be applied to other astrophysical scenarios for the spectral-siren method, as well as for other techniques to understand the robustness of the standard siren

methods in inferring the cosmological parameters, primarily the Hubble parameter $H(z)$. Apart from systematic errors due to astrophysical populations of BBHs, the impact of a few other sources of systematic errors in GW cosmology, such as inclination angle (M. Müller et al. 2024; A. Salvarese & H.-Y. Chen 2024), inaccurate waveform (N. Kunert et al. 2024), peculiar velocity of host galaxy (S. Mukherjee et al. 2021a; H. Nimmonkar & S. Mukherjee 2023), and photometric redshifts uncertainty of host galaxy (C. Turski et al. 2023), were explored previously.

This paper is organized as follows. We outline the setup of the Blinded-MDC in Section 2. The simulated mock data for different astrophysical cases considered in this analysis are discussed in Section 3. The formalism we have used for the analysis is discussed in Section 4. The results and the discussion for different cases from the Blinded-MDC are presented in Section 5. Finally, the conclusion and future scopes are discussed in Section 6.

2. Setup for the Blinded Mock Data Challenge

The Blinded-MDC is set up to find out the robustness of a GW cosmology analysis pipeline in inferring the cosmological parameters in two aspects, namely (i) in the presence of statistical uncertainties of the GW source parameters without any modeling error in the astrophysical population of BBHs and (ii) in the presence of modeling error(s). For this Blinded-MDC, we have focused on two scenarios—(i) the vanilla case: where the underlying population model is POWER LAW + GAUSSIAN PEAK (PLG) without any redshift evolution (R. Abbott et al. 2023a), but the values of the parameters in PLG (discussed in the next Section, 3.1) and the values of the cosmological parameters are blinded, and (ii) the redshift-dependent scenario: where the underlying population model is PLG with redshift-dependent parameters (as discussed in Section 3.2). This differs from the usual assumption of the no redshift evolution of the mass distribution of BBHs (B. P. Abbott et al. 2021; R. Gray et al. 2023; S. Mastrogianni et al. 2024).

The injected values are chosen from a wide prior range of astrophysical and cosmological parameters and the Blinded-MDC is made blind to ensure that the choice in the analysis settings is independent of the injected values, to test the reliability of the pipelines, and to avoid any confirmation bias. As a result, two disjoint research groups are prepared: (i) the simulation team and (ii) the analysis team, which are coordinated by the MDC coordinator. The exact setup of the Blinded-MDC is shown by a schematic diagram in Figure 1. We describe below each of the cases.

Blinded-MDC Simulation setup: The code GWSIM (C. Karathanasis et al. 2023b) is used to generate the simulated GW mock data using the LVK O1+O2+O3+O4 noise power spectral densities (PSDs) for the vanilla case and with the O4 noise PSD for the redshift-dependent scenario (Y. Aso et al. 2013; LIGO Scientific Collaboration et al. 2015; F. Acernese et al. 2015) with the duty cycles as mentioned in Table 1. The values of the BBH source parameters, merger rate parameters, and cosmological parameters are drawn randomly from a fixed prior range before passing them to the simulation code GWSIM. The GWSIM code has gone through an internal review for the validation of the injected simulation set with the expected distribution for the model of the astrophysical population. For the vanilla case, the simulation set that was passed to the analysis team is for the values for which the GWSIM code was

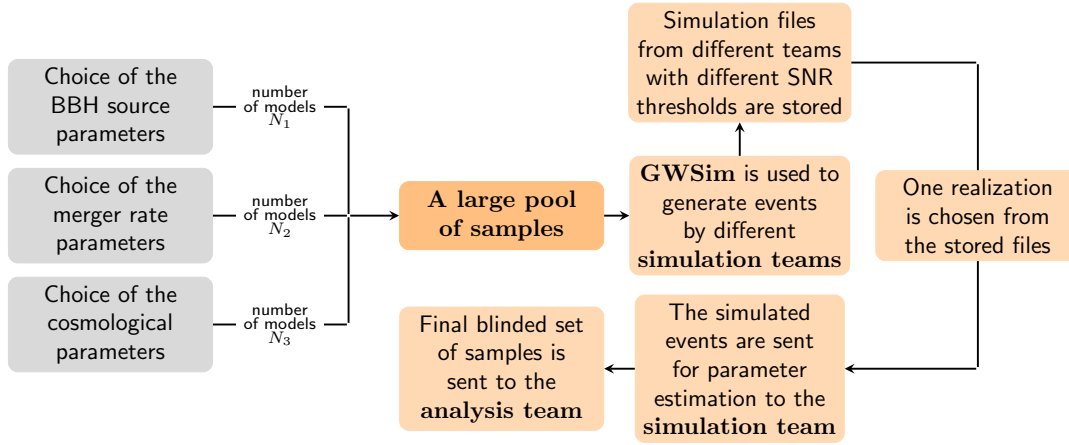


Figure 1. Schematic representation of the Blinded-MDC setup.

Table 1

List of Detectors with the Duty Cycle Used in the Analysis

For the Vanilla Case with $T_{\text{obs}} = 3$ yr.				
Detector	Observation Run			
	O1	O2	O3	O4
H	0.6	0.6	0.75	0.75
L	0.5	0.6	0.75	0.75
V	...	...	0.75	0.75
For the redshift-dependent scenario with $T_{\text{obs}} = 1$ yr.				
Detector	Observation Run			
	O1	O2	O3	O4
H	...	...	...	0.75
L	...	...	...	0.75
V	...	...	...	0.75

Note. The sensitivity considered for the analysis corresponds to the relative observing run.

validated. This is to ensure that there is no error in the pipeline of Blinded-MDC setup, which includes both injection and analysis parts. The values of the injected parameters and whether they are the same as those for which the code is tested remained blinded until the analysis was completed. The matched filtering signal-to-noise ratio (SNR) denoted by ρ thresholds used for the vanilla case is $\rho_{\text{th}} = 10$. For the redshift-dependent scenario, the matched filtering SNR threshold is $\rho_{\text{th}} = 12$. Since the redshift-dependent MDC scenario explores the impact on H_0 due to the mismodeling of the astrophysical population assumption, we have chosen a higher SNR threshold for this case to make sure that the contamination from noise fluctuation is not significant, and hence impact from statistical fluctuation is limited on the error budget of the cosmological and astrophysical population. This helps in understanding more clearly the effect of systematic errors over the statistical uncertainties.

The injection set is prepared for 10 cases (denoted by $I(j, \rho_{\text{th}})$, where the index j runs from $j = 1$ to $j = 10$) for each threshold with the values of the cosmological and astrophysical population parameters drawn randomly from a uniform

distribution over the range by different members of the simulation team (at least by three members). Then, a particular simulation from $I(j', \rho_{\text{th}})$ is randomly chosen by the coordinator and a simulation team member for which estimation of the GW source parameters is performed using the package Bilby (G. Ashton et al. 2019). The particular realization (denoted by j') chosen for the PE and the corresponding values of the cosmological and astrophysical parameters chosen are not known to anyone in the Blinded-MDC team until the end of the analysis by the analysis team. As for every different realization, the underlying population parameters and the cosmological parameters are different, a randomly chosen realization makes the injection part blind to everyone, except the coordinator and one member of the simulation team. For the other simulations, Bilby parameter estimations were not performed to reduce the computation cost. However, these sets are available after the unblinding and can be used for any checking purposes.

Blinded-MDC analysis setup: The code *icarogw* (S. Mastrogiovanni et al. 2024) is used for the analysis purpose of exploring the interplay between the cosmological parameters and the population parameters. The analysis team used a realization given by the simulation team of the posteriors on the GW source parameters ($\mathcal{M}$, q , d_L , decl., R.A., i , Ψ , ϕ , and t_c) without any information about the cosmological and astrophysical population parameters used to generate them. Using this set, the analysis team performs the joint inference of the cosmological parameters and the population parameters for the underlying model considered as the fiducial astrophysical model PLG without redshift evolution and Flat Λ CDM cosmological model. The analysis team explores the impact of different choices such as the number of samples in the selection function (or in the injection set), the impact of wider priors, and the impact on the inference with a change in the number of detected events. Finally, once the analysis members are satisfied with all the checks, the values of the cosmological and the population parameters used for the simulations are unblinded and compared with the inferred values.

The vanilla case allows us to verify the analysis setup on whether the analysis code can infer the values correctly for a known scenario. For the second case, i.e., redshift-dependent scenario, the Blinded-MDC inspects the impact on the inference of the parameters (both astrophysical and cosmological), when the underlying true astrophysical model differs from the fiducial assumption of the model. In the remainder of

this paper, we describe in detail the results from the injection and the analysis parts.

3. Cases Studied in the Blinded-MDC

As introduced in Section 2, the scope of the Blinded-MDC is twofold, in the first place we want to quantify statistical uncertainties under the assumption that the settings used to generate and analyze data are the same. This is the vanilla case, which allows us to assess the impact due to detector noise and the fact that our data represent a limited sample. Specifically, we want to study the presence of statistical fluctuations and outliers among data that might bias the overall results. Second, our goal is to evaluate systematic errors arising from astrophysical mismodeling. To do this, we will only assume a redshift dependence in the data generation process, not in the analysis. We address this in the redshift-dependent scenario. We provide a detailed explanation of the astrophysical models used to create the events data set for both scenarios.

In this work, the GW source parameter estimation for the simulated mock events is performed with Bilby (G. Ashton et al. 2019) using the IMPRphenomPv2 waveform approximant (S. Husa et al. 2016; M. Hannam et al. 2014), using the standard Bilby priors for BBHs with all the spin parameters fixed to zero (delta-function prior). An example event analysis is shown in the Appendix in Figure 21. The analysis is done on the following nine parameters: the chirp mass $\mathcal{M}$ (a combination of the two-component masses in detector frame²³), the mass ratio q (defined as the ratio between the lightest and the heaviest mass component), the luminosity distance d_L to the source, the sky position given by the R.A. and decl., the inclination angle ι (defined as the angle between the line of sight and the perpendicular to the orbital plane of the binary), the polarization angle Ψ , the phase ϕ of the GW signal and the time of coalescence t_c . The choice of the nonspinning waveform will not have any noticeable impact on the cosmology results obtained in this paper for this detector noise.

3.1. Vanilla Model

We assume black holes to have an astrophysical origin and consequently model the binary merger rate $R(z)$ with a function similar to the one used for fitting the star formation

by the parameters γ and κ as

$$R(z) = R_0(1+z)^\gamma \frac{1 + (1+z_p)^{-(\gamma+\kappa)}}{1 + \left(\frac{1+z}{1+z_p}\right)^{(\gamma+\kappa)}}, \quad (1)$$

where R_0 is the merger rate of BBHs at $z = 0$ and z_p denotes the peak of the merger rate. The simulated events are drawn from the PLG source-frame probability distribution, motivated by the analysis of previously detected GWs events (R. Abbott et al. 2023a; C. Karathanasis et al. 2023a), which can be analytically expressed by

$$P_1(m_1|m_{\min}, m_{\max}, \alpha) = \begin{cases} (1-\lambda)\mathcal{P}(m_1, -\alpha) + \lambda G(m_1, \mu, \sigma), & m_{\min} < m_1 < m_{\max}, \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

$$P_2(m_2|m_{\min}, m_1, \beta) = \begin{cases} \mathcal{P}(m_2, -\beta), & m_{\min} < m_2 < m_1, \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where m_1 and m_2 are source-frame primary and secondary mass, respectively. $\mathcal{P}$ and G are normalized power-law (PL) function and Gaussian function, respectively. For this case, 286 GW events are detected with an SNR threshold of $\rho_{\text{th}} = 10$ for three years of observation time and duty cycle as given in Table 1. The sampled events detector-frame mass distributions are shown in Figures 2(b)–(c). The mock event distribution as a function of distance is shown in Figure 2(a). The population and rate parameters used to obtain the events are given in Table 2.

3.2. Redshift-dependent Scenario

In this scenario, we make the same assumptions as the vanilla model, but the mass model parameters now include an additional linear dependence on redshift. In this case, we obtained events with an SNR greater than an SNR threshold $\rho_{\text{th}} = 12$ with one year of observation time and duty cycle as given in Table 1. The total number of events selected is 80 for this case. The primary mass and secondary mass distribution is modeled with all parameters of the mass distributions varying with redshift as

$$P_1(m_1|m_{\min}, m_{\max}, \alpha, z) = \begin{cases} ((1-\lambda(z))\mathcal{P}(m_1, -\alpha(z)) + \lambda(z) G(m_1, \mu(z), \sigma(z)), & m_{\min}(z) < m_1 < m_{\max}(z), \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

rate (SFR; P. Madau & M. Dickinson 2014). We use a parametric form for the BBHs merger rate evolution captured

$$P_2(m_2|m_{\min}, m_1, \beta, z) = \begin{cases} \mathcal{P}(m_2, -\beta(z)), & m_{\min}(z) < m_2 < m_1(z), \\ 0, & \text{otherwise}, \end{cases} \quad (5)$$

We model the redshift dependence of each parameter using a linear relation of the form $x(z) = x_0 + z \epsilon_x$, where x_0 denotes the parameter value at redshift $z = 0$ (C. Karathanasis et al. 2023b). This form of redshift evolution is applied uniformly across all mass distribution parameters in the redshift-

²³ The chirp mass is defined as in terms of component masses (m_1, m_2):

$$\mathcal{M} \equiv \frac{(m_1 m_2)^{\frac{3}{5}}}{(m_1 + m_2)^{\frac{1}{5}}}$$

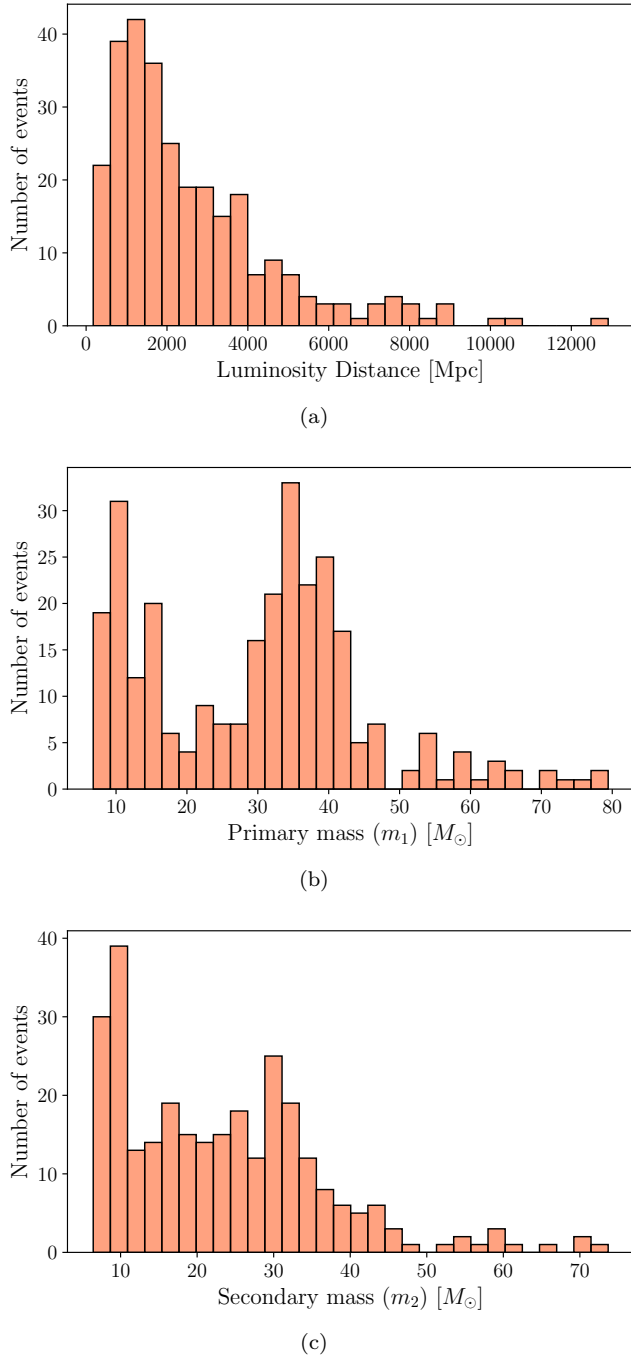


Figure 2. Histogram of the number of detected events as a function of (a) luminosity distance, (b) detector-frame primary mass, and (c) detector-frame secondary mass for the vanilla scenario.

dependent scenario. The specific injected values for these parameters are summarized in Table 2, while the prior distributions used for sampling are detailed in Table 3. The high value of the Hubble constant chosen in the blinded setup checks for the robustness of the analysis and scrutinizes whether there is any sort of confirmation bias in the inference of the Hubble constant. We demonstrate later that this large value of the Hubble constant is not driving any bias in the inference.

The motivation for introducing redshift dependence arises from the fact that stellar evolution is significantly influenced by the environment in which stars form. Studies show that in

low-metallicity environments, more massive stars are more likely to form compared to high-metallicity environments (G. Dopcke et al. 2013; J. Li et al. 2023). This occurs because low-metallicity gases are less efficient at cooling, leading to a higher Jeans mass and reducing the likelihood of fragmentation into smaller stars (C. J. Clarke & V. Bromm 2003; G. Dopcke et al. 2013). Additionally, stars with higher metallicity tend to lose more mass through stellar winds compared to their low-metallicity counterparts, resulting in lighter remnants (J. S. Vink et al. 2001; J. T. van Loon 2006; M. Mokiemi et al. 2007). Consequently, Black Holes (BHs) are expected to be more massive at high redshifts, where metallicity is lower compared to low redshifts. The impact of metallicity evolution on BBHs population from GWTC-3 was previously studied, indicating mild hints toward redshift-dependent mass distribution (S. Mukherjee 2022; C. Karathanasis et al. 2023a). However, no conclusive evidence of it is not found from GWTC-3 (T. A. Callister 2024).

The primary mass distribution for the redshift-dependent scenario is illustrated in Figure 3. The peak (μ_g) of the GAUSSIAN component of the primary mass distribution is located at $40.9 M_\odot$ at $z = 0$, shifting to higher masses at a rate of $2.84 M_\odot$ per unit redshift. Similarly, the standard deviation of the GAUSSIAN (σ_g) is $7.63 M_\odot$ at $z = 0$, increasing linearly with redshift at a rate of $2.7 M_\odot$. The PL index of the primary mass distribution takes a value of 5.3 at $z = 0$, increasing linearly with redshift at a rate of 2.99, while the PL index of the secondary mass distribution decreases at a rate of 1.063 per unit redshift.

For the merger rate, we assume a Madau–Dickinson-like redshift evolution (P. Madau & M. Dickinson 2014), similar to that in a redshift-independent vanilla scenario (see Equation (1)). The merger rate for the injected parameters is relatively flat (see Figure 4), with a maximum value at $z = 0$ before gradually decreasing with increasing redshift. The maximum redshift of the injection is 1.7. The sources are located at relatively smaller luminosity distances compared to the vanilla scenario, primarily due to the high injected value of $H_0 = 173.28 \text{ km s}^{-1} \text{ Mpc}^{-1}$. The higher detection threshold (ρ_{th}) also contributes to selecting nearer sources, but its impact on the overall distance distribution is subdominant. It is important to note that the redshift dependence in the simulated mass distribution is implemented directly in redshift space. Therefore, the compression in the luminosity distance distribution caused by the high H_0 does not influence the effect of the redshift evolution in the mass distribution.

In Figures 5(a), (b), and (c), we depict the histogram of luminosity distance (d_L) and the component masses (m_1 , and m_2), respectively, of the injected sources. The minimum mass is set to $7.8 M_\odot$ at $z = 0$, with only a tiny redshift dependence. Compared to the vanilla scenario, there is a relatively smaller fraction of sources between $10 M_\odot$ and $30 M_\odot$, which is a consequence of the steeper PL for both the primary and secondary mass distributions. The luminosity distance of the sources in this scenario is more concentrated at lower values compared to the vanilla scenario. This is because the merger rate here is relatively constant with redshift, whereas in the vanilla scenario it increases steeply with redshift.

4. Methods

In what follows, we first give an overview of the spectral-sirens method used to infer at the same time cosmological and

Table 2
Summary of Injected Values for the Two Scenarios: Vanilla and Redshift-dependent

Parameter	Description	Injected Value	
		Vanilla	Redshift-dependent
ΛCDM cosmological model			
H_0	Hubble constant in [km s ^{−1} Mpc ^{−1}]	67.8	173.28
$\Omega_{m,0}$	Matter energy density today	0.3	0.3
Power Law + Gaussian peak mass model			
α	PL index of primary mass	3.4	$5.325 + 2.99z$
β	PL index of secondary mass	0.8	$3.05 - 1.063z$
m_{min}	Minimum source mass in $M_\odot$	5	$7.8 + 0.324z$
m_{max}	Maximum source mass in $M_\odot$	100	$87.58 + 17.16z$
δ_m	Smoothing factor in $M_\odot$ at low-mass cut-off	4.8	3.19
μ_g	Peak of the Gaussian in $M_\odot$	35	$40.9 + 2.84z$
σ_g	Sigma of the Gaussian in $M_\odot$	3.9	$7.63 + 2.70z$
λ_{peak}	Fraction of events in Gaussian in [0, 1] interval	0.04	$0.04 + 0.43z$
Madau-Dickinson rate model			
γ	PL exponent of rate ($z \lesssim z_p$)	2.7	0.03
κ	(Negative of) PL exponent of rate ($z \gtrsim z_p$)	2.9	2.92
z_p	Rate parameter (turnover point)	1.9	3.86
R_0	Local merger rate in [Gpc ^{−3} yr ^{−1}]	20	16.1

Note. We report a description for each of the parameters in the cosmological assumptions (Λ CDM) with different H_0 values, the mass model (POWER LAW + GAUSSIAN PEAK, with redshift dependence in the second scenario), and the Madau–Dickinson rate model with different parameters.

Table 3
Prior Ranges used to Sample the Injection Parameters for the Redshift-dependent Scenario

Parameter	Prior Used for Sampling
H_0	$\mathcal{U}(10, 200)$
$\Omega_{m,0}$	0.3 (Fixed)
α_0	$\mathcal{U}(1.5, 12)$
β_0	$\mathcal{U}(-4, 12)$
$m_{\min}$	$\mathcal{U}(2, 10)$
$m_{\max}$	$\mathcal{U}(70, 200)$
δ_m	$\mathcal{U}(0, 10)$
μ_g	$\mathcal{U}(10, 70)$
σ_g	$\mathcal{U}(0.4, 10)$
λ_{peak}	$\mathcal{U}(0, 1)$
ϵ_α	$\mathcal{U}\left(-\frac{\alpha}{z_{\max}}, \frac{\alpha}{z_{\max}}\right)$
ϵ_β	$\mathcal{U}\left(-\frac{\beta}{z_{\max}}, \frac{\beta}{z_{\max}}\right)$
$\epsilon_{m_{\min}}$	$\mathcal{U}\left(0, \frac{m_{\min}}{z_{\max}}\right)$
$\epsilon_{m_{\max}}$	$\mathcal{U}\left(0, \frac{m_{\max}}{z_{\max}}\right)$
ϵ_{μ_g}	$\mathcal{U}\left(\max\left(0, \epsilon_{m_{\min}} + \frac{m_{\min} - \mu_g}{z_{\max}}\right), \min\left(\frac{\mu_g}{z_{\max}}, \epsilon_{m_{\max}} + \frac{m_{\max} - \mu_g}{z_{\max}}\right)\right)$
ϵ_{σ_g}	$\mathcal{U}\left(0, \frac{\sigma_g}{z_{\max}}\right)$
ϵ_λ	$\mathcal{U}\left(-\frac{\lambda}{z_{\max}}, \frac{1 - \lambda}{z_{\max}}\right)$
γ	$\mathcal{U}(0, 12)$
κ	$\mathcal{U}(0, 6)$
z_p	$\mathcal{U}(0, 4)$
R_0	$\mathcal{U}(0, 100)$

population parameters given the two starting sets of BBH events: one for the vanilla case and one for the redshift-dependent scenario. We present a reconstruction of both the

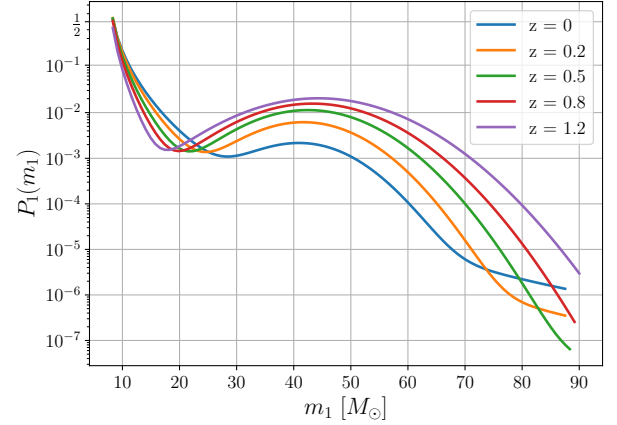


Figure 3. Primary mass distribution for the redshift-dependent scenario at different redshifts.

rate and mass models for two cases, comparing them to the distributions used to sample the two populations. The main results on the inference of H_0 are then presented and discussed in full details.²⁴

4.1. Spectral-sirens Method: *icarogw* Setup

We use the *icarogw* package (S. Mastrogiorganni et al. 2024) to infer the population and cosmological parameters simultaneously (a brief description of all the parameters is provided in Table 2). Within this package, we use Bilby’s (G. Ashton et al. 2019) sampler *dynesty* (J. S. Speagle 2020) to generate posterior samples for the hyperparameters. The source distributions assumed in the analysis coincide with the ones used during the generation of the GW population only in

²⁴ All the scripts used to produce the results in this paper are available on GitHub repository (https://git.ligo.org/users/sign_in).

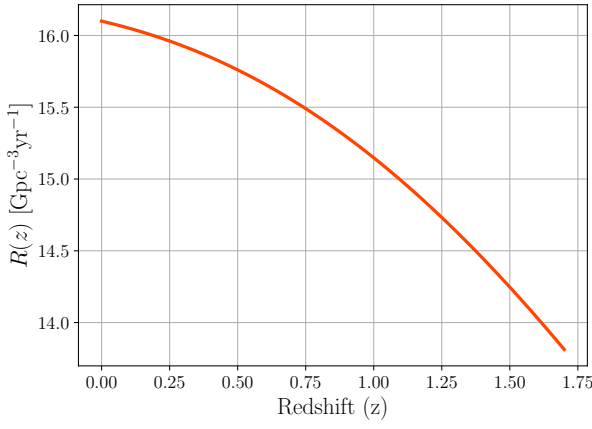


Figure 4. Merger rate as a function of redshift for the redshift-dependent scenario with the fiducial values mentioned in Table 2.

the vanilla case. For the redshift-dependent scenario, we keep the same assumptions of the vanilla analysis, even though the underlying population is now redshift-dependent.

The combined posteriors on the source population and cosmological hyperparameters, given N_{obs} GW detections each with data $\{x\} = (x_1, x_2, \dots, x_{\text{obs}})$, are obtained using the following equation (I. Mandel et al. 2019; S. Vitale et al. 2022; S. Mastrogiovanni et al. 2021):

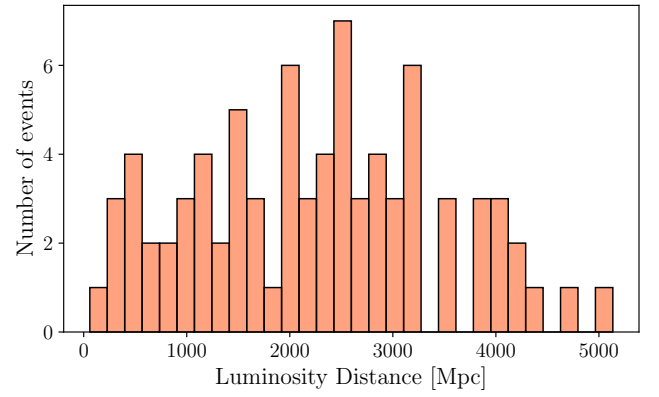
$$p(\Lambda|\{x\}, N_{\text{obs}}) \propto \pi(\Lambda) \prod_{i=1}^{N_{\text{obs}}} \frac{\int p(x_i|\Lambda, \theta) p_{\text{pop}}(\theta|\Lambda) d\theta}{\int p_{\text{det}}(\theta, \Lambda) p_{\text{pop}}(\theta|\Lambda) d\theta}, \quad (6)$$

where $\pi(\Lambda)$ is the prior on the hyperparameters (both population parameters and cosmological ones, see Table 4). Since each GW event is independent, we multiply across the N_{obs} observations. The probability of each GW event can be broken down into two main components: the parameters of the individual source, represented by θ , and the population-induced prior, p_{pop} , which describes the expected distribution of the hyperparameters at the population level.

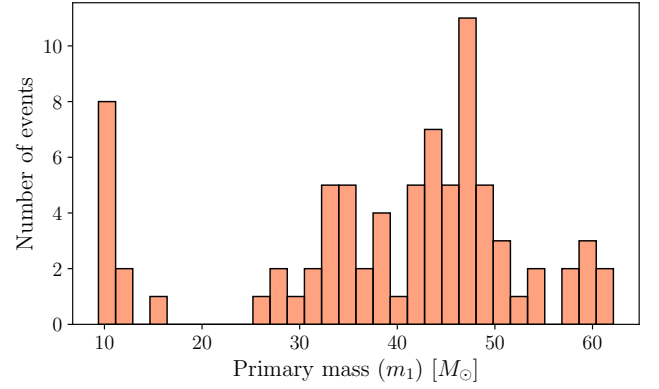
The denominator of the likelihood, $\int p_{\text{det}} p_{\text{pop}} d\theta$, corrects for the selection bias. The data set that is analyzed has been obtained after applying selection criteria. In this study, we required that the events' match-filtering SNRs ρ are above a fixed threshold ($\rho_{\text{th}} = 10$ for the vanilla case and $\rho_{\text{th}} = 12$ for the redshift-dependent scenario). This selection must be accounted for to avoid the Malmquist bias (K. G. Malmquist 1922; T. J. Loredo 2004; I. Mandel et al. 2019). The correction method consists of estimating the probability of detection of GW events, using the very same selection criteria and for any values of hyperparameters Λ , describing the cosmology, the population parameters, and the mergers rate (see Table 2). This probability of detection is given by the ratio of the expected number of detected events $N_{\text{exp}}(\Lambda)$ to the total number of mergers $N(\Lambda)$:

$$\frac{N_{\text{exp}}(\Lambda)}{N(\Lambda)} = \int p_{\text{det}}(\theta, \Lambda) p_{\text{pop}}(\theta|\Lambda) d\theta, \quad (7)$$

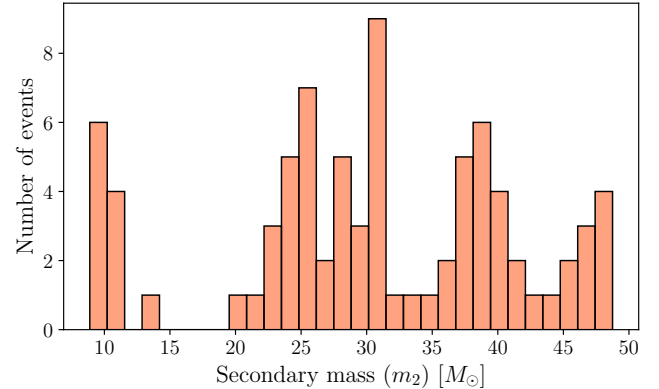
θ being the set of the individual parameters of the binary system: $\theta = (m_1, m_2, z, \iota, \text{R.A., decl.} \dots)$. We chose to study a population model that only depends on the mass and redshift so that we restrict θ to $\theta = (m_1, m_2, z)$ and we consider the



(a)



(b)



(c)

Figure 5. Histogram of the number of detected events as a function of (a) luminosity distance, (b) detector-frame primary mass, and (c) detector-frame secondary mass for redshift-dependent scenario.

domain where the events are detected, i.e., when $p_{\text{det}} = 1$:

$$\frac{N_{\text{exp}}(\Lambda)}{N(\Lambda)} = \int_{\rho \geq \rho_{\text{th}}} p_{\text{pop}}(m_{1s}, m_{2s}, z|\Lambda) dm_1 dm_2 dz. \quad (8)$$

The last integral is estimated by Monte Carlo integration using a large set of N_{sim} realizations $(m_{1,i}, m_{2,i}, z_i)$ corresponding to N_{det} detected ($\rho \geq \rho_{\text{th}}$) simulated events. These injected events are randomly drawn from an initial probability density function $\pi(m_1, m_2, z|\Lambda)$ so that the probability of detection is

Table 4

List of Priors used in the Analysis for Both the Vanilla and the Redshift-dependent Scenarios

	Priors	
	Vanilla	Redshift-dependent
H_0	$\mathcal{U}(30, 140)$	$\mathcal{U}(10, 250)$
$\Omega_{m,0}$	$\mathcal{U}(0.1, 0.9)$	$\mathcal{U}(0.1, 0.9)$
α	$\mathcal{U}(1.5, 12)$	$\mathcal{U}(1.5, 12)$
β	$\mathcal{U}(-4, 12)$	$\mathcal{U}(-4, 12)$
$m_{\min}$	$\mathcal{U}(2, 10)$	$\mathcal{U}(2, 10)$
$m_{\max}$	$\mathcal{U}(50, 200)$	$\mathcal{U}(50, 200)$
δ_m	$\mathcal{U}(0, 10)$	$\mathcal{U}(0, 10)$
μ_g	$\mathcal{U}(20, 50)$	$\mathcal{U}(10, 80)$
σ_g	$\mathcal{U}(0.4, 10)$	$\mathcal{U}(0.4, 10)$
λ_{peak}	$\mathcal{U}(0, 1)$	$\mathcal{U}(0, 1)$
γ	$\mathcal{U}(0, 12)$	$\mathcal{U}(0, 12)$
κ	$\mathcal{U}(0, 6)$	$\mathcal{U}(0, 6)$
z_p	$\mathcal{U}(0, 4)$	$\mathcal{U}(0, 4)$
R_0	$\log \mathcal{U}(10^{-2}, 10^3)$	$\log \mathcal{U}(10^{-2}, 10^3)$

evaluated as:

$$\begin{aligned} \frac{N_{\text{exp}}(\Lambda)}{N(\Lambda)} &= \int_{\rho \geq \rho_{\text{th}}} p_{\text{pop}}(m_1, m_2, z|\Lambda) dm_1 dm_2 dz, \\ &= \frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{det}}} \frac{p(m_{1,i}, m_{2,i}, z_i|\Lambda)}{\pi(m_{1,i}, m_{2,i}, z_i|\Lambda)}. \end{aligned} \quad (9)$$

4.2. Priors Settings

We examine 14 hyperparameters listed in Table 2. We assign broad and uninformative priors to each of them. The priors for the hyperparameters in the two scenarios are summarized in Table 4. The prior ranges of the three parameters differ between the two scenarios: H_0 , μ_g , and σ_g . Initially, the prior ranges for both scenarios were the same. After our preliminary results exhibited railing on either side of their prior range, we decided to enlarge the prior settings for the redshift-dependent scenario. We refer to this aspect of the study as the impact from astrophysical prior on H_0 inference in the remaining paper. Due to the blinded nature of the MDC (as discussed in Section 2), different prior choices are made for the analysis.

5. Results from the Blinded-MDC

5.1. Vanilla Model Results

The results of the vanilla model serve as an overall validation of the setup and demonstrate its consistency with the initial assumptions. We start from the reconstruction of both the rate and mass models, as shown in Figures 6 and 7. The colored lines represent the posterior samples from the 68% highest density interval, while the solid black lines indicate the starting distribution derived from the injection values in Table 2. We find good agreement with the injected rate and mass distributions.

Our estimate for H_0 using all 286 events and the full set of 10^6 injections is $62.08^{+28.07}_{-21.15} \text{ km s}^{-1} \text{ Mpc}^{-1}$ at 1σ , compared to the injected value of $67.8 \text{ km s}^{-1} \text{ Mpc}^{-1}$. In most cases, the inferred values of the parameters are recovered at the 1σ level,

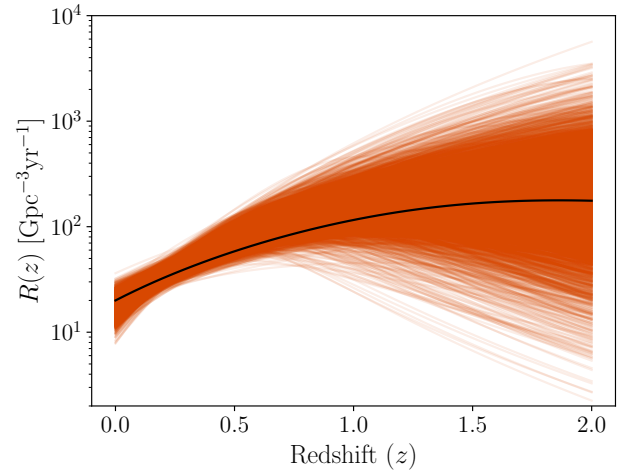


Figure 6. Posterior sample plots for the vanilla scenario following Equation (1). The black line represents the injected values. The colored lines show the posterior samples from the 68% highest density interval.

with some exceptions, such as α , $m_{\min}$, σ_g , and γ , which are slightly outside the 1σ range but within 2σ , as can be seen from the full corner plot in Figure 23. Figure 8 illustrates a reduced corner plot that highlights the distributions of the H_0 posterior samples along with the parameters that demonstrate the strongest correlations with H_0 , specifically μ_g and σ_g , which are negatively correlated with H_0 . This means that higher values of μ_g suggest a more massive BBH population, leading to lower inferred redshifts and a lower H_0 to match observed signals at specific luminosity distances.

The three different colored regions of the 2D posterior in all the corner plots shown in this paper represent, from the darkest to the lighter shade, respectively, the 39.3%, 86.5%, and 98.9% C.I. contours for a 2D Gaussian distribution. The dashed lines in the corresponding 1D histograms instead show the 68% C.I. of the marginalized distribution.

We then performed the following series of checks to pinpoint potential sources of statistical uncertainties:

1. *Impact of the injection samples: number and realization.* To study the robustness of the population inference scheme, we vary the realization of GW signals, (referred to as injection samples), which are used to compute the denominator of Equation (6). The denominator is evaluated as a Monte Carlo sum, hence we expect to have more accurate results for a larger amount of samples. The hyperparameter posterior is obtained for three simulations, with 10^4 , 10^5 , and 10^6 samples. We generally find that the inferred values of all hyperparameters are compatible with their true values at 90% credible region. In Figure 9, we show H_0 posteriors for the three cases. We also varied the sample of 10^5 injections and obtained comparable results, as shown in Figure 10. The distribution confirms that the evaluation of the denominator partially depends on the number of injections and not the specific realization of the sample, once we reached a sufficient number of injections for the denominator to converge, which in our case means 10^5 injections. The posteriors for all the parameters are shown in Appendix C in Figure 26 varying the number of injections and in Figure 27 varying the injection sample.
2. *Dependence on the number of events.* We study how the inferred cosmological and population parameters depend

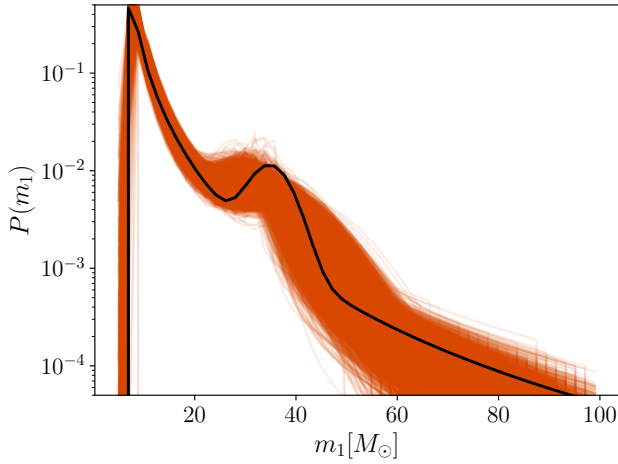


Figure 7. Posterior sample plots for the vanilla scenario for the primary mass distribution following Equation (2). The black line represents the injected values. The colored lines show the posterior samples from the 68% highest density interval.

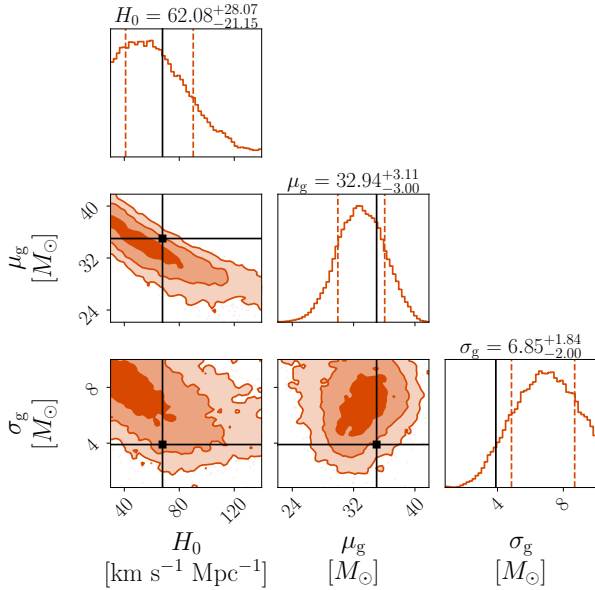


Figure 8. Reduced corner plot for the vanilla scenario of H_0 and parameters that mostly correlate with H_0 , which are μ_g and σ_g . The injected (true) values are plotted as black lines. The three different colored regions of the 2D posterior represent, from the darkest to the lighter shade, respectively, the 39.3%, 86.5%, and 98.9% C.I. contours for a 2D Gaussian distribution. The dashed lines in the corresponding 1D histograms instead show the 68% C.I. of the marginalized distribution. For the full corner plot refer to Figure 23.

on the number of GW events used for the analysis. We compare the results using different numbers of GW events, specifically [150, 180, 200, 220, 250, 286], with 10^5 injections (once we assessed 10^5 and 10^6 injections were producing the same outcome, as shown in the previous point). The results shown in Figure 11 demonstrate that we obtain more accurate constraints on the parameters as more GW events are used. In Figure 11, we show H_0 posteriors only for clarity (the full parameter set is displayed in Appendix C in Figure 28).

3. *Dependence on event sample realization.* Due to the homogeneity principle, the analysis of subsets of the observed GW catalog is expected to yield parameter

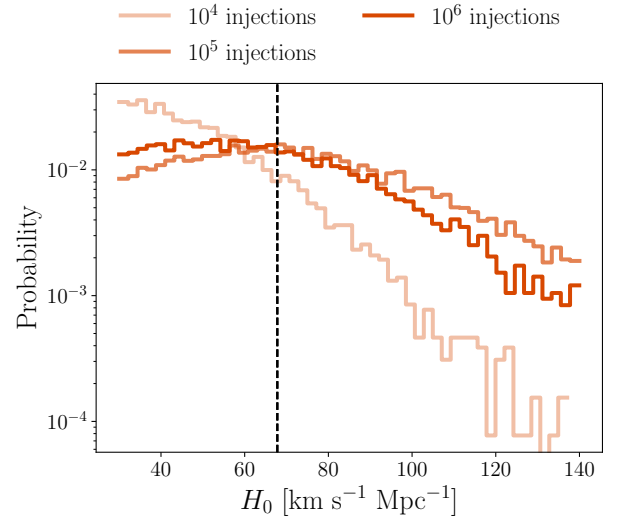


Figure 9. Check with the vanilla scenario events. Posterior samples of the H_0 parameter varying the number of injections. The dashed black line is the injected value. See also the full plot with all the parameters in Figure 9.

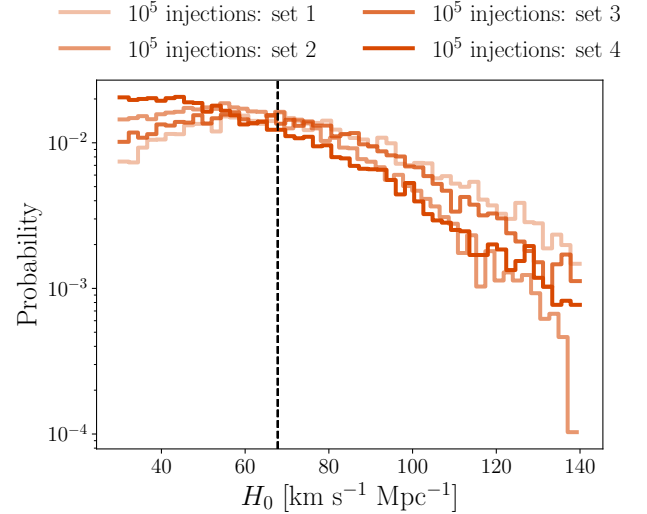


Figure 10. Check with the vanilla scenario events. Posterior samples of the H_0 parameter varying the realization sample with 10^5 injections. The dashed black line is the injected value. See also the full plot with all the parameters in Figure 27.

values with random deviations (which statistically average out) from the parameter values extracted from the whole catalog. For further validation of the analysis scripts, we test this expectation. We split the observed GW catalog in batches of 50 and analyze the resulting subpopulations with 10^5 injections used to compute the denominator of Equation (6). In Figure 12, we show some example results for the posterior distribution restricted to H_0 . We find that there are no outliers in our sets of events. The posteriors are summarized in Appendix C in Figure 29.

In summary, the Blinded-MDC analysis for the vanilla scenario shows that when the underlying true model of the astrophysical population is the same as the model used in the analysis, the inference of parameters is not biased. However, due to the degeneracy between the cosmological parameters and the astrophysical population parameters, the inferred value

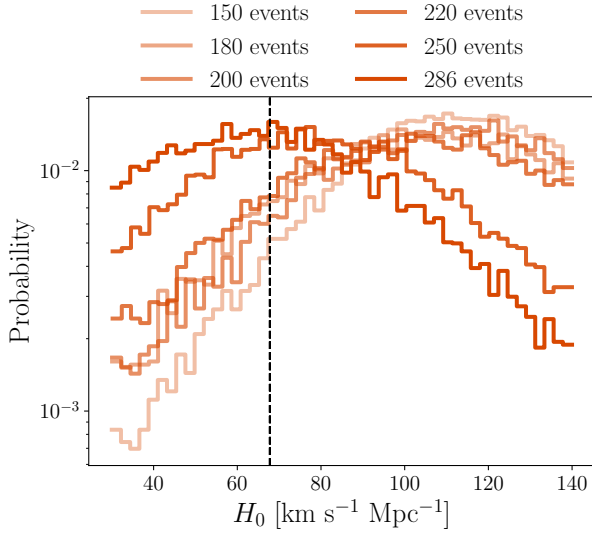


Figure 11. Check with the vanilla scenario events. Posterior samples for the H_0 parameter using a different number of events for the vanilla case. The dashed black line is the injected value. See also the full plot with all the parameters in Figure 28.

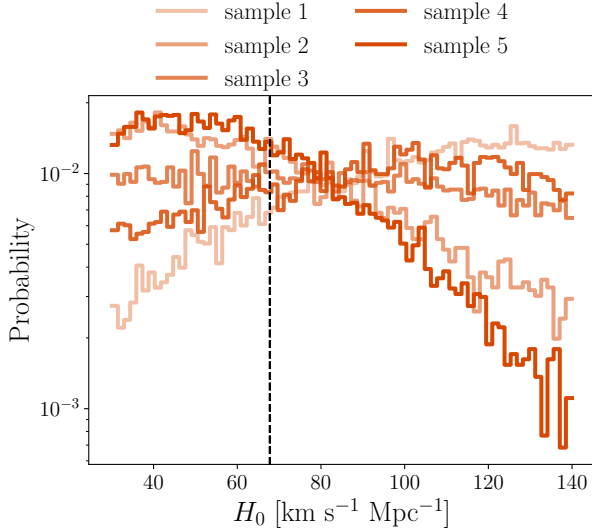


Figure 12. Check with the vanilla scenario events. Posterior samples of the H_0 parameter varying the events sample. The dashed black line is the injected value. See also the full plot with all the parameters in Figure 29.

of H_0 can shift away from the true value. We also show the impact of prior on the astrophysical parameters in inferring H_0 in the repository. The plot in Figure 30 indicates that wider prior on the parameters related to black hole mass distribution moves the H_0 posterior to a lower value. The vanilla model analysis served as a validation tool to assess the robustness of the inference pipeline against statistical and sampling uncertainties under the scenarios analyzed above: injection sample size and realization, number of events, and event sample realization.

5.2. Redshift-dependent Scenario Results

The goal of analyzing the redshift-dependent scenario is to understand the impact of the redshift evolution of the mass distribution on the inference of the Hubble constant due to mismodeling the astrophysical population model of BBHs by assuming a redshift-independent scenario. So, we use the

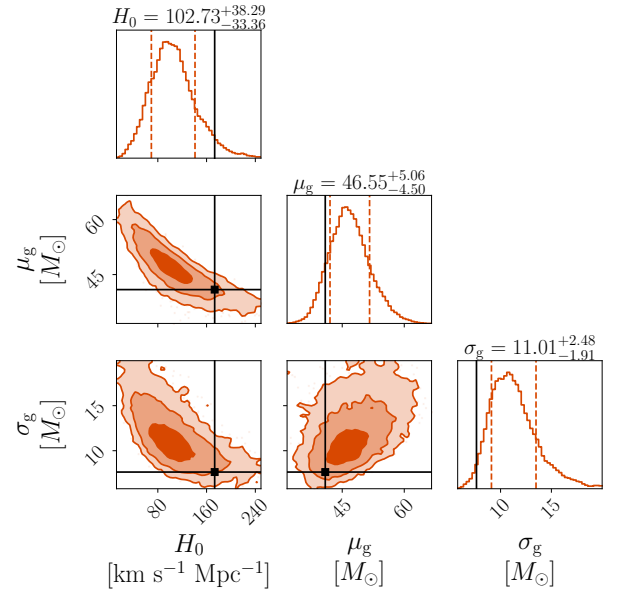


Figure 13. Reduced corner plot for the redshift-dependent scenario of H_0 and parameters that mostly correlate with H_0 , which are μ_g and σ_g . The injected (true) values are plotted as black lines. The three different colored regions of the 2D posterior represent, from the darkest to the lighter shade, respectively, the 39.3%, 86.5%, and 98.9% C.I. contours for a 2D Gaussian distribution. The dashed lines in the corresponding 1D histograms instead show the 68% C. I. of the marginalized distribution. For the full corner plot refer to Figure 24. See also Figures 14 and 15 for a redshift evolution of μ_g and σ_g .

redshift-independent PLG model, which is currently a setup for the population inference, for the analysis of these mock simulations. However, it is important to stress here that though the analysis is performed for a specific mass model, the impact on the inference of H_0 can happen for any other astrophysical population model as well if there is a mismatch with the simulation.

We perform a joint parameter estimation for the astrophysical population and cosmological parameters using *icarogw* on 80 detected events in the mock data. Our main result for the redshift-dependent scenario is the one obtained with all 80 events and the full set of 10^6 injections. The corresponding full corner plot with the parameter posteriors is shown in the Appendix in Figure 24. The main result we find is that the injected value of Hubble constant is outside $\sim 90\%$ Credible Interval (C.I.) of the inferred posterior distribution. A few other parameters related to the mass distribution of the GW sources such as α , β , $m_{\min}$, μ_g , σ_g , λ_{peak} show about 1σ discrepancy from the injected value.

To explore the impact of redshift evolution of the mass distribution on the value of H_0 , we show the joint contour of H_0 , μ_g , and σ_g in Figure 13, and the comparison between the injected value of μ_g and σ_g with the recovered posterior from all the events in Figures 14 and 15, respectively. These plots indicate that the posterior on the parameters $\mu_{g,z}$ and σ_z agrees with the injected values at different redshifts (shown by color bar) within about 68% C.I. of the inferred posterior distribution. Furthermore, the samples within 68% of the mass distribution for different redshift bins of the injected value are shown in Figure 16 along with the injected distribution. We divided the population into three sets, redshift-wise (we chose $z < 0.65$, $0.65 \leq z < 1.05$, and $z \geq 1.05$ to ensure a similar number of events in each bin). We then compare the reconstructed posteriors with three reference models, evaluated

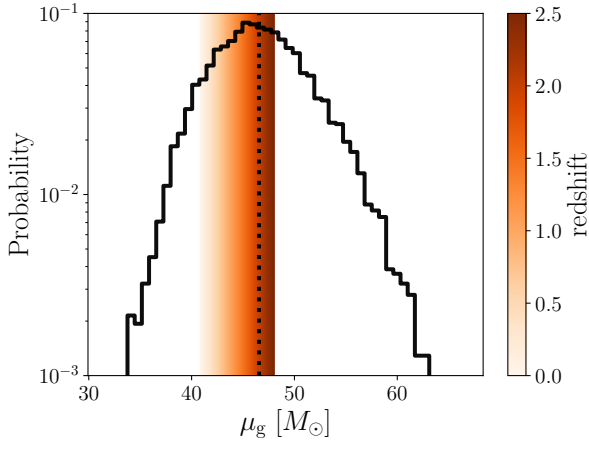


Figure 14. Posterior samples for μ_g for the redshift-dependent scenario plotted in black. The dotted vertical line represents the median of the distribution. The colored region represents the values of the injected μ_g , $40.9 + 2.84z$, at different redshifts specified in the color bar.

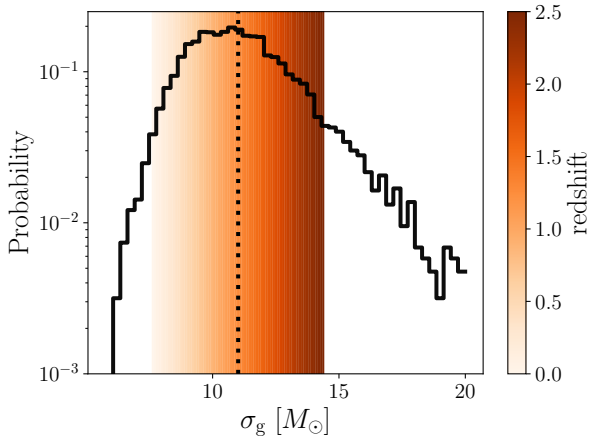


Figure 15. Posterior samples for σ_g for the redshift-dependent scenario plotted in black. The dotted vertical line represents the median of the distribution. The colored region represents the values of the injected σ_g , $7.63 + 2.70z$, at different redshifts specified in the color bar.

at parameter values in the center of each redshift bin following Table 2. In all cases, we used posterior samples within the 68% highest density interval. These show a good agreement with the injected values in the simulation.

Among the parameters that control the merger rate, such as R_0 , γ , and κ , the recovered values match the injected value well because the underlying model is the same and there is no negative impact on the inference of the Hubble constant. The parameter z_p which denotes the peak of the BBH merger rate distribution does not show a good recovery due to the fewer number of detected events at the injected value of high redshift ($z_p = 3.86$), as shown in the full corner plot in Figure 24 (given in the Appendix). The merger rate realizations from the samples of the posteriors with the 68% C.I. are shown in Figure 17.

In Figure 18, we show the impact on H_0 from varying the number of events included in the analysis, specifically [35, 45, 55, 65, 75, 80]. The distribution shows nearly consistent behavior between all the subsamples having different number of events. This indicates that the statistical uncertainty due to a fewer number of events is not causing any

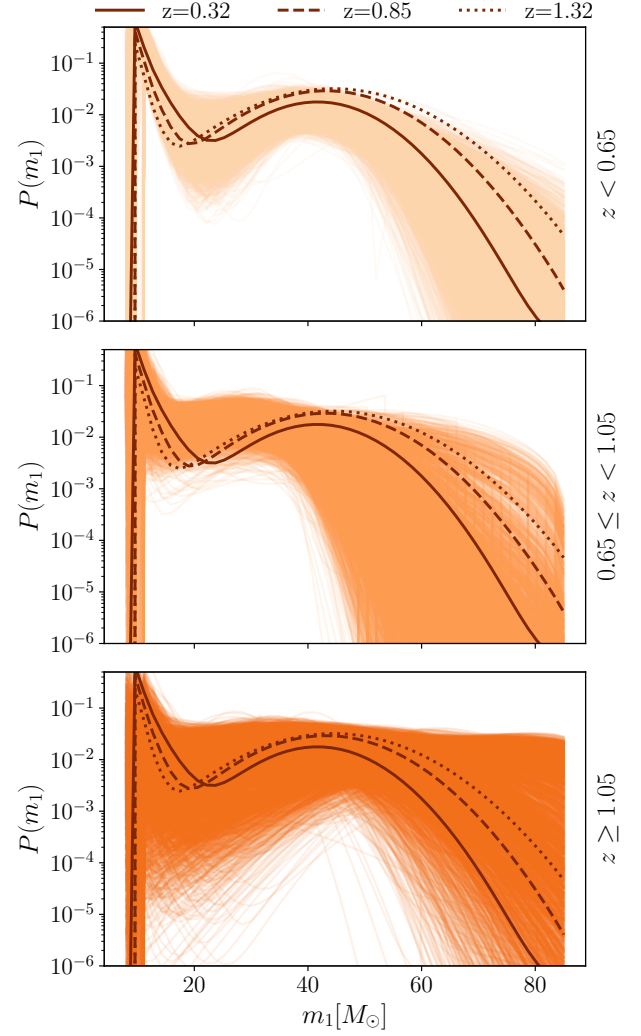


Figure 16. Posterior sample plots for the redshift-dependent scenario for the primary mass distribution following Equation (4). In each plot, the colored lines show the posterior samples from the 68% highest density interval. These samples have been categorized based on the redshift range of the events used for the analysis. The redshift range is shown on the left y-axis: The top panel represents events with $z < 0.65$, the central panel shows events with $0.65 \leq z < 1.05$, and the bottom panel corresponds to events with $z \geq 1.05$. In each subplot, there are three darker lines (solid, dashed, and dotted) that remain consistent and represent the model at different redshift values, as defined in Table 2. These redshift values have been selected to fall in the middle of the redshift ranges we divided the events into. The highest-redshift event is at $z = 1.7$.

pronounced difference in the observed discrepancy in the H_0 posterior from the injected value shown by the black dashed line. For all the cases, the peak of the inferred H_0 posterior is at least 76 percentile away (toward a lower side) from the injected value.

We also performed statistical tests on the impact of the injection samples and dependence on event sample realization. For the former, we used 10^5 and 10^6 injection set samples, and for the latter we split the GW events into batches of 25 to analyze the resulting subpopulations. (For all statistical tests, the results obtained are similar to the cases obtained for the vanilla case.)²⁵ The impact of prior on the astrophysical parameters in inferring H_0 is also shown in the repository. The

²⁵ The plots for these tests are shown in the repository (https://git.ligo.org/users/sign_in).

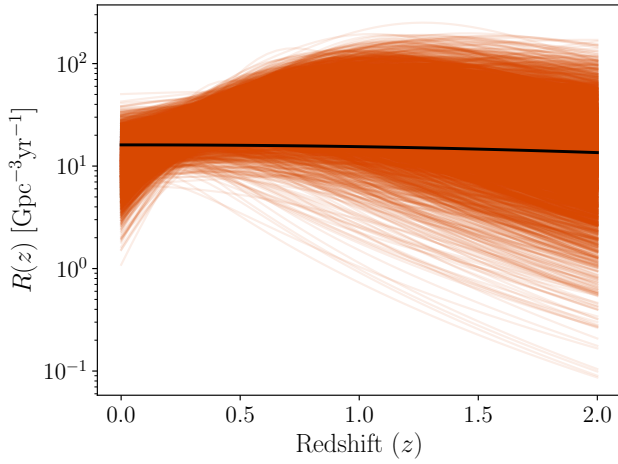


Figure 17. Same as in Figure 6 but for the redshift-dependent scenario.

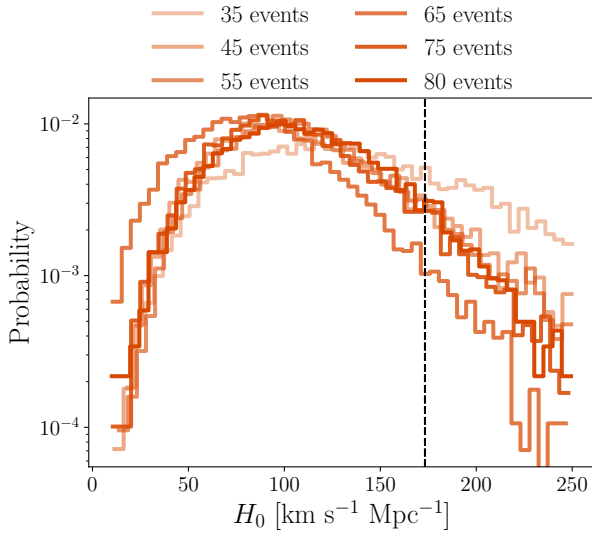


Figure 18. Check with the redshift-dependent scenario events. Posterior samples for the H_0 parameter using a different number of events. The dashed black line is the injected value.

conclusion is similar to the vanilla case, i.e., the change in the prior mass distribution has a noticeable impact on H_0 posterior.

5.2.1. Discussion

In this section, we discuss our findings from the Blinded-MDC for the redshift-dependent scenario. As stated at the beginning of the motivation section, the reason for exploring a redshift-dependent mass model is to understand its impact on the inference of cosmology, due to ignorance of such evolution in reality due to complicated BBH formation channels and its dependence on parent star metallicity. As a result, we consider a mock simulation with redshift evolution in the mass and merger rate, where the merger rate evolution model agrees with the analysis model, but the mass distribution in the mock data differs from the one used in the analysis. The corresponding results obtained from the Blinded-MDC were presented in the previous section.

The key result we are interested in exploring is to understand what is the impact of the mismodeling on the inference of the Hubble constant. We find that the value of the Hubble constant is outside the 90% C.I. from the injected value. The difference

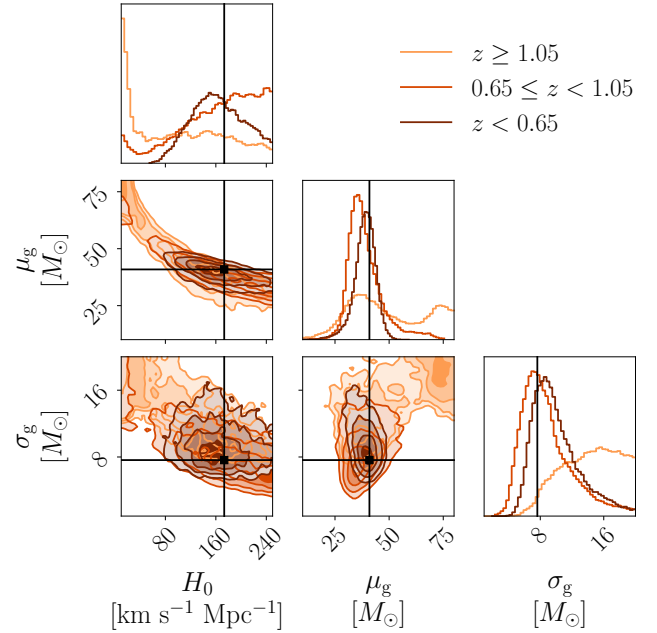


Figure 19. Reduced corner plot for H_0 , μ_g , and σ_g parameters for the redshift-dependent scenario. We show results obtained selecting events in three different redshift ranges—at low redshift $z < 0.65$, medium $0.65 \leq z < 1.05$, and at redshifts higher than $z \geq 1.05$. The maximum redshift does not exceed $z = 1.7$. The injected values are shown in black. For μ_g and σ_g they refer to the injected value at $z = 0$ as in Table 2. The three different colored regions of the 2D posterior represent, from the darkest to the lighter shade, respectively, the 39.3%, 86.5%, and 98.9% C.I. contours for a 2D Gaussian distribution. The dashed lines in the corresponding 1D histograms instead show the 68% C.I. of the marginalized distribution.

between the peak of the posterior distribution from the injected value is primarily driven by the correlation between the Hubble constant with the parameters μ_g and σ_g . The higher value of the inferred μ_g and σ_g is driving a lower value of the Hubble constant. On comparing Figures 8 and 13, we find that the incorrect population assumption can drive the inference of the μ_g and σ_g toward a higher value, resulting into a shift in the Hubble constant posterior from the injected value.

To understand the origin of this discrepancy better, we scrutinize the results with events in three groups, divided based on their source redshift (which are known from the simulation set) as group 1 ($z \in \{0, 0.65\}$), group 2 ($z \in \{0.65, 1.05\}$), and group 3 ($z > 1.05$) with a nearly equal number of sources in each group. The analysis in groups is performed by implementing the product over samples in Equation (6) in these three groups. The selection function of the analysis does not change for this subdivided run, except for a reduction in T_{obs} by one-third for each group. It can be seen from Equation (6) that the term in the denominator, which captures the selection function, does not depend on the index (denoting individual events) over which the product in the numerator is carried out. The corresponding results on H_0 (after marginalizing over other parameters) are shown in Figure 19. For the results from group 1, the value of H_0 matches very well with the injected value of the Hubble constant, as shown by the solid black line. This is also valid for the other parameters μ_g and σ_g which are strongly correlated with H_0 . However, as we move toward higher redshift bins, i.e., group 2 and group 3, the posterior on H_0 moves toward a lower value, and the posteriors on μ_g and σ_g moves toward a higher value. The inference of the value of Hubble constant consistent with the injected value

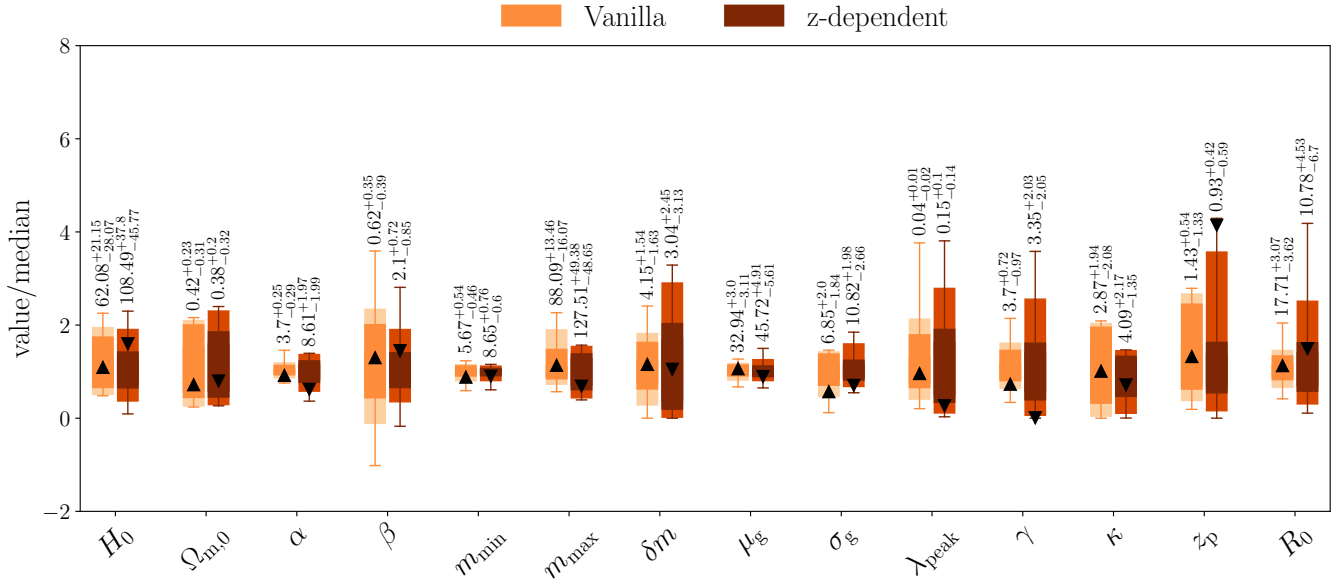


Figure 20. Summary plot of results for both the vanilla (lighter shade) and the redshift-dependent (darker shade) models. The runs used all the available events (286 for the vanilla case and 80 for the redshift-dependent one) with 10^6 injections for selection effects. The x -axis represents the inferred hyperparameters, while the y -axis shows all the posterior samples divided by the median value of each distribution for easier comparison on a single plot. The boxplots are designed to encompass 1σ of the data around the median in the darker shade and up to 2σ with the lighter shade. The whiskers extend from the minimum to the maximum value to represent all the data. The black triangles represent the injected values divided by the corresponding median of the distribution. The text represents the median value of the parameter and the 68% confidence interval around it. For the redshift-dependent model, the values shown in Table 2 (considering $z = 0$) are being referred to.

from the low redshift bin confirms that when there is only little redshift evolution of the mass distribution, the inference of the Hubble constant is correct and there is no confirmation bias. However, when the mass distribution is redshift evolving, the posterior distribution of the Hubble constant drives away from the injected value.

This happens because the sources contributing from the lowest redshift bin (in group 1) have the least impact on the redshift evolution of the mass distribution for the injection model discussed in Section 3.2. As a result, when one is making an inference using a model with no redshift evolution, the error due to mismodeling is minimal. On the other hand, for sources in group 2 and group 3, the redshift evolution of the μ_g and σ_g parameter moves the values by about 10% and 50%, respectively, toward a higher value in comparison to the lowest redshift bin case, resulting in larger mismodeling. As a result, the inferred value of the Hubble constant H_0 is shifted toward a lower value from the injected one, as shown in Figure 19. The shift toward a lower value arises due to a strong negative correlation with the mass parameters (μ_g , σ_g) and H_0 . Since the values at a higher redshift have intrinsically higher mass distribution, an analysis model that does not capture this effect associates this to a higher value of μ_g and σ_g , resulting in a lower value of inferred redshift than the true one.²⁶ As a result, the corresponding inferred H_0 is lower than the true value of the H_0 used in the simulation. This effect gets pronounced from group 2 to group 3 samples, since the changes due to redshift in μ_g and σ_g are large.

To exclude that the lower ρ_{th} of the vanilla case gives better constraints on the recovered hyperparameters due to a larger number of events, we ran a vanilla analysis imposing a $\rho_{th} = 12$ as in the redshift-dependent scenario. The results, shown

compared to the $\rho_{th} = 10$ ones in Figure 25, confirm the ones with the lower threshold, showing slightly enlarged error bars, as a consequence of a reduced number of events (164 when $\rho_{th} = 12$). This check sets the vanilla and the redshift-dependent analysis on the same footing, and corroborates our findings on the presence of a bias in the redshift-dependent scenario.

The comparison of the marginalized posterior with respect to the injected value does not show any strong deviation for μ_g and σ_g as shown in Figure 20 due to large uncertainties. However, the value of H_0 starts showing deviation at 91 percentile from the injected value due to the reasons mentioned above. This implies that even though the lowest distance bin value of H_0 does not show any discrepancy, a mild 10% variation in μ_g can lead to noticeable deviation from the injected value due to the sources at higher distance. Though due to the blinded nature of the Blinded-MDC, in this analysis the value of the injected Hubble constant H_0 is large, the key finding showing the impact of redshift evolution of mass distribution will cause a discrepancy in the inferred value of H_0 remains valid.

Along with the impact on the Hubble constant H_0 , we also find that the inference of the BBH merger rate parameters such as γ and z_p also get impacted with a deviation from the injected value at about 2σ or more, when the underlying population model of the mass distribution is the redshift-dependent scenario, and it is not taken into account in the analysis. Among these two parameters, the z_p parameters show more deviation, as can be seen in Figure 24. The constraints on the parameter z_p are usually weak and show a long tail due to the impact of selection function, which limits the detection of a large number of sources from high redshifts. The parameters controlling the mass distribution are inferred within 1.5σ from the injected value for both the vanilla case and redshift-dependent scenario, with a maximum deviation noticed for the parameters α and σ_g , as shown in Figure 24.

A redshift evolution of the black hole mass distribution and also the merger rate is an expected outcome from an

²⁶ Since the observed masses m_z are redshifted, $m_z = m(1+z)$. An incorrect inference of true source-frame mass will lead to an incorrect inference of redshift z by m_z/m .

Table 5
Summary of the Systematic and Statistical Tests Performed in the Analysis

Type of Tests	Objective	Vanilla Scenario	Redshift-dependent Scenario
Systematic Uncertainties	Inference of H_0	Not impacted	Impacted
	Impact of population prior on H_0	Impacted	Impacted
	Inference of BBH mass distribution	Mildly impacted	Mildly impacted
	Inference of BBH merger rate	Not impacted	Impacted
Statistical Uncertainties	Impact of number of samples in p_{det}	Impacted	Impacted
	Impact of subsample of events	Mildly impacted	Mildly impacted

astrophysical perspective due to the dependencies of these on progenitor metallicities and SFRs. The metallicity of the Universe varies by a few orders of magnitude with redshift, with a decreasing trend at high redshift. Furthermore, from galaxy to galaxy, there is a large variation in the metallicity. As a result, along with a monotonic redshift evolution, there is going to be additional stochastic behavior in the masses of BHs at a redshift. This can lead to a dispersive mass distribution from any monotonic behavior in redshift. In this Blinded-MDC, we have considered a simplistic scenario of linear evolution with redshift of the parameters controlling the mass distribution (as discussed in Section 3.2). However, in reality, the mass distribution evolution can be even more severe. Future analysis will be conducted to understand this impact for a larger number of GW samples and better sensitivity for several astrophysical scenarios.

The matter energy density $\Omega_{m,0}$ is the other cosmological parameter inferred in our analyses. However, all of our runs (see Figures 23 and 24) show practically no constraining power on $\Omega_{m,0}$.

6. Conclusion

This work presents the first Blinded-MDC for the inference of the cosmological parameters from GW observations feasible using BBHs from the currently ongoing network of LVK detectors with O4 sensitivity (LIGO Scientific Collaboration et al. 2015; F. Acernese et al. 2015; B. P. Abbott et al. 2018). This study aims to explore the reliability of standard siren cosmology in inferring the true values of the cosmological parameters when the dark siren technique using a BBH mass distribution is used assuming an underlying model. Through this Blinded-MDC, we explore both statistical and systematic uncertainty due to the interplay between astrophysical model assumptions on the BBH mass distribution and the inference of Hubble constant H_0 from the GW data.

To explore this aspect, we considered two astrophysical scenarios, namely (i) the vanilla model, which represents the currently known model of BBH mass distribution, and (ii) a redshift-dependent scenario that considers a mass distribution model with redshift evolution, which differs from the fiducial assumption. For both cases, GW mock samples are generated from the code GWSIM (C. Karathanasis et al. 2023b) by two independent simulation teams which are disjointed from the analysis team members. The key findings from this analysis are listed below and also mentioned in the Table 5:

1. For a given astrophysical model consistent with the vanilla scenario, the cosmology inference code can infer the injected value for all the parameters correctly in a blinded setup using the mass distribution (see Figure 8).
2. The impact of fewer samples in the injection set for the calculation of the detection probability mildly

- overestimates (underestimates) the posterior on the Hubble constant H_0 at the lower (higher) values in comparison to a case with more samples (see Figure 9).
3. If the mass distribution of the BBHs shows mild redshift evolution (as shown in Figure 3) which is theoretically unknown (or unknown from independent observations) due to the evolution of stellar properties with redshift, then the inferred value of the Hubble constant shows discrepancy from the injected value at more than 1σ for O4 noise sensitivity (see Figure 13).
4. For the case of redshift evolving mass distribution, the observed statistical fluctuation in the inferred value of the parameters (in particular the value of H_0) is not significant in comparison to the systematic error due to mass mismodeling (see Figure 18).
5. The inferred parameters describing the astrophysical mass distribution do not show any significant departure from the injected parameters of the mass distribution when the underlying true model is unknown. As a result, even if the mass distribution evolves mildly (considered in this analysis), the values of μ_g and σ_g agree well with the inferred posterior distribution (see Figures 14 and 15).
6. The last point implies that in an analysis performed in different redshift bins of the GW sources, the mass distribution in individual redshift bins does not show any significant deviation from the injected value due to large error bars. However, the combined posterior on the Hubble constant leads to a noticeable discrepancy from the true value (See Figure 19).
7. For the redshift-dependent case, the inference of some of the parameters which control the merger rate, namely γ and z_p , gets partially impacted when the underlying population model is unknown. We find that both of these parameters are impacted by around 2σ or more. However, it needs to be pointed out that the value of z_p is poorly constrained, also due to the selection effect arising from the detection of fewer sources at high redshift.

This study indicates that the analysis method and setup of the code work well when the model used in the mock data generation is the same as the one used in the analysis. But if there are departures in the mass model that are redshift dependent, then even though the impact on the inference of the astrophysical population model is not pronounced, it becomes more pronounced on the inference of H_0 . This arises because the value of H_0 gets constrained by all the sources located at different redshift ranges. If there is a monotonic evolution (increase or decrease) in the mass distribution of the BBHs, then the inferred redshift starts to differ from the true value of the redshift in a monotonic way. As a result, the corresponding value of the Hubble constant H_0 will

monotonically decrease or increase. In reality, inferred mass distribution with a wrong model may still give a reasonable astrophysical understanding of the mass distribution but a value of H_0 toward a lower (or higher) value from the true value, depending on whether the mass distribution is increasing (or decreasing) with redshift. The findings from the Blinded-MDC obtained here have a crucial impact on the robustness of standard siren cosmology when the astrophysical assumptions of mass distribution are made. As expected, the discrepancy gets pronounced with a higher number of GW events.

In the future, it will be crucial to understand the interplay between the astrophysical assumptions and cosmological parameters with better sensitivity and a higher number of events. It will be important to investigate the robustness in inferring cosmological parameters for different astrophysical scenarios of BBH formation, depending on their delay time distribution, and stellar metallicity. Along with these, an understanding of the robustness of other dark sirens techniques related to galaxy incompleteness and the assumptions on host galaxy luminosity for the statistical host techniques will be necessary (A. G. Hanselman et al. 2025; G. Perna et al. 2025). Similarly, for the cross-correlation technique, an understanding of the impact on H_0 inference due to the redshift dependence of the GW bias parameter will also be necessary (C. C. Diaz & S. Mukherjee 2022; A. Dehghani et al. 2025). In summary, this study performs the first Blinded-MDC and provides a framework to check the reliability of cosmological inferences from the standard sirens cosmology for the scenarios of dark sirens, which can be extended for other techniques as well. To make a robust measurement of H_0 from standard sirens which can shed light on the ongoing H_0 -TENSION (L. Verde et al. 2019; E. Abdalla et al. 2022), Blinded-MDC of the analysis techniques will be crucial to explore possible sources of modeled and unmodeled systematic uncertainties. We will explore different such scenarios in future Blinded-MDC analyses.

Acknowledgments

The authors are very thankful to Archisman Ghosh for reviewing the manuscript as a part of the LIGO Publication and Presentation policy and providing valuable comments on the draft. The authors also thank to the LIGO-Virgo-KAGRA Scientific Collaboration for providing noise curves. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants PHY-0757058 and PHY-0823459 and the computing resources of the `<data|theory>` Universe-Lab, supported by TIFR and the Department of Atomic Energy, Government of India. LIGO is funded by the U.S. National Science Foundation (NSF) and Virgo is supported by the French CNRS, Italian INFN, and Dutch Nikhef, along with contributions from Polish and Hungarian institutes. This collaborative effort is backed by the NSF's LIGO Laboratory, a major facility fully funded by the National Science Foundation. This research leverages data and software from the Gravitational Wave Open Science Center, a service provided by LIGO Laboratory, the LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA. Advanced LIGO's construction and operation receive support from STFC of the UK, Max-Planck Society (MPS), and the State of Niedersachsen/Germany, with additional backing from the Australian Research Council. Virgo, affiliated with the European Gravitational Observatory (EGO), secures funding through contributions from various European institutions. Meanwhile,

KAGRA's construction and operation are funded by MEXT, JSPS, NRF, MSIT, AS, and MoST. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. This work is supported in part by the Perimeter Institute for Theoretical Physics. Research at Perimeter Institute is supported 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 Economic Development, Job Creation, and Trade. M.C.E. is supported by COSMOGRAV. The research of F.B. and C.T. is supported by Ghent University Special Research Funds (BOF) project BOF/STA/202009/040, the interuniversity iBOF project BOF20/IBF/124, and the Fonds Wetenschappelijk Onderzoek (FWO) research project G0A5E24N. They also acknowledge support from the FWO International Research Infrastructure (IRI) grant I002123N for Virgo collaboration membership and travel to collaboration meetings. The work of S.M. and M.R.S. is part of the `<data|theory>` Universe-Lab, supported by TIFR and the Department of Atomic Energy, Government of India. A.E.R. was supported by the UDEA projects 2021-44670, 2019-28270, and 2023-63330. R.G. is supported by STFC grant ST/V005634/1. This project has received financial support from the CNRS through the AMORCE funding framework and from the Agence Nationale de la Recherche (ANR) through the MRSEI project ANR-24-MRS1-0009-01, Spanish Research Project PID2021-123012NB-C43 [MICINN-FEDER], and Centro de Excelencia Severo Ochoa Program CEX2020-001007-S at IFT.

Appendix A

Bilby Runs for Events PE

We used 286 events and 80 BBH events for the vanilla and the redshift-dependent scenarios, respectively. The population models used to generate the events are described in Section 3. We employed Bilby (G. Ashton et al. 2019) to obtain the parameter estimation for all of these events. The events analysis settings of the two models are common and are described in the following.

As stated in the main text, the analysis was done for the $\mathcal{M}$, q , d_L , decl., R.A., i , Ψ , ϕ , and t_c parameters (see a brief description in Section 3). The remaining spin parameters (six in total) were fixed to zero with a δ -prior. We used the `dynesty` sampler and the `IMRPhenomPv2` waveform approximant. This waveform does not include higher-order harmonics. While it does incorporate spin precession effects, we did not vary the spin parameters in our analysis. Although both higher-order modes and spin precession can help break degeneracies between parameters, their impact on detection and cosmological inference is expected to be limited given the current detector sensitivity. Therefore, we adopted this simplified waveform setup to reduce computational complexity. The detectors settings are outlined in Table 1. For the vanilla scenario, we considered 3 yr of observations corresponding to a combination of the first four observing runs. Instead, we limited our analysis to the fourth observing run corresponding to 1 yr of observing time for the redshift-dependent case.

In Figures 21 and 22, we show the parameter estimation results for the vanilla and the redshift-dependent scenarios, respectively. The posterior samples for the nine parameters considered in the analysis are shown in orange and the true (injected) values are indicated with a black line. We chose one event per scenario as a reference example.

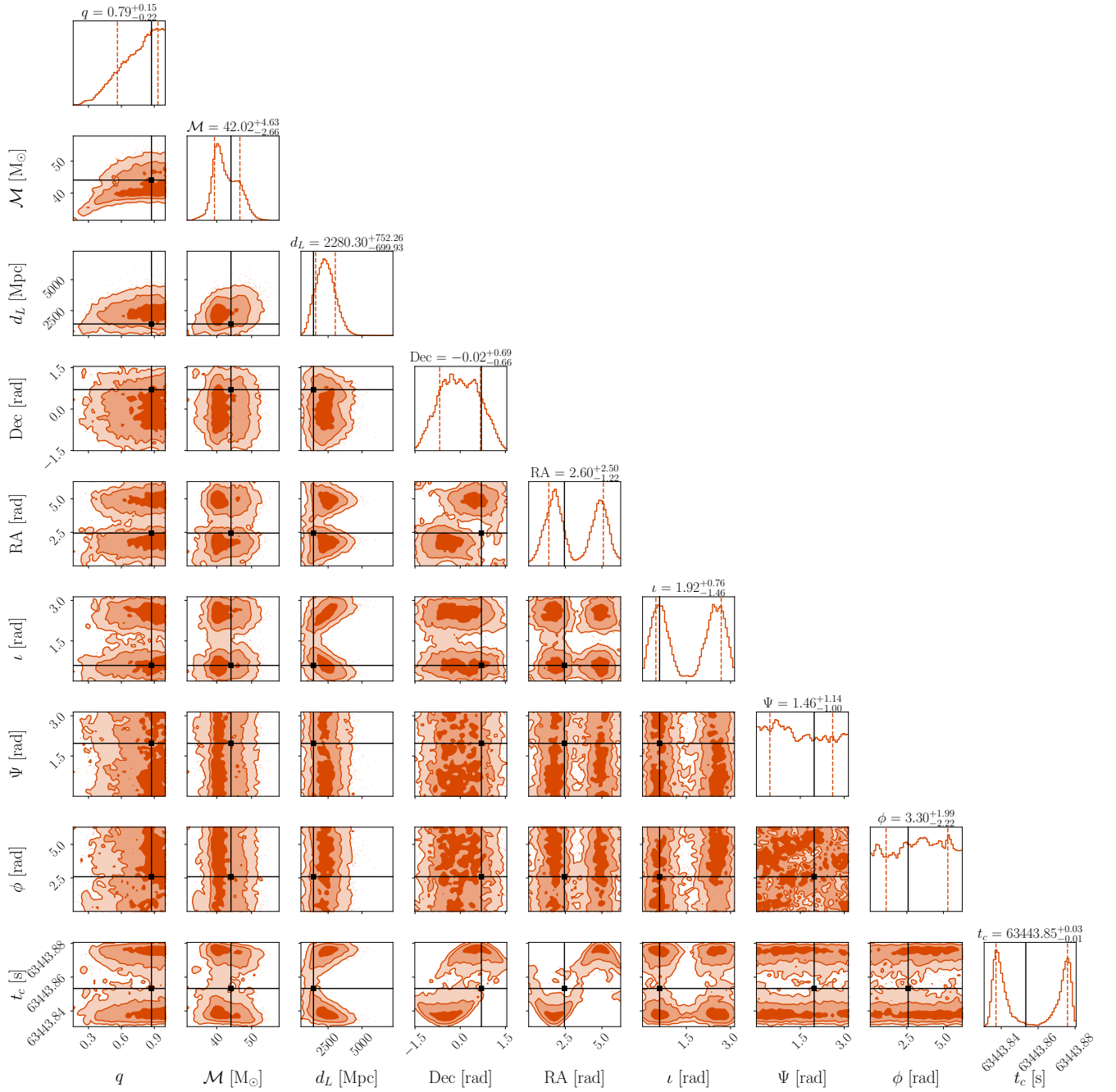


Figure 21. Posterior samples from the Bilby run for one event from the vanilla case. The injected values are indicated in black. The three different colored regions of the 2D posterior represent, from the darkest to the lightest shade, respectively, the 39.3%, 86.5%, and 98.9% C.I. contours for a 2D Gaussian distribution. The dashed lines in the corresponding 1D histograms instead show the 68% C.I. of the marginalized distribution.

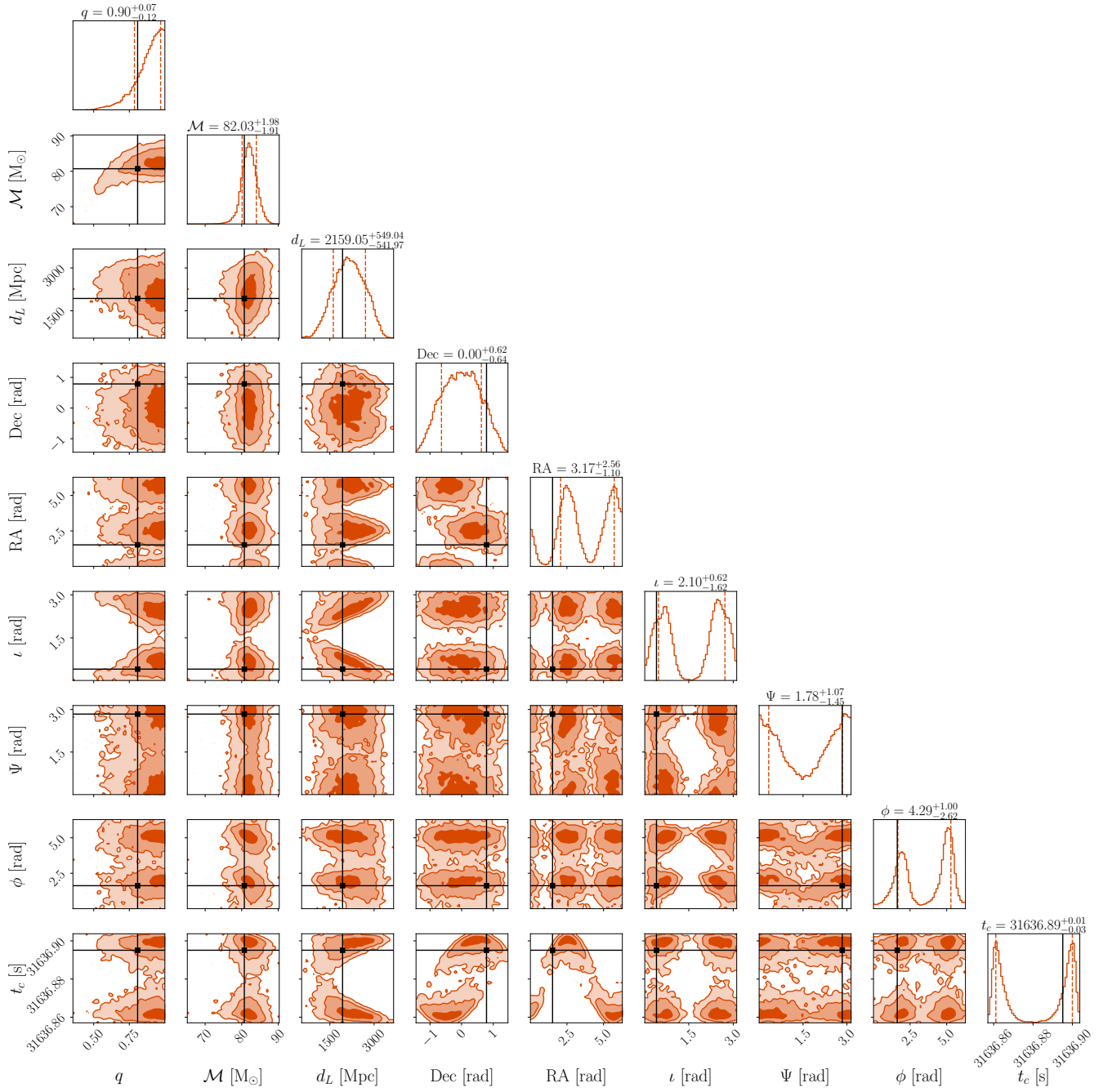


Figure 22. Posterior samples from the `Bilby` run for one event from the redshift-dependent scenario. The injected values are indicated in black. The three different colored regions of the 2D posterior represent, from the darkest to the lightest shade, respectively, the 39.3%, 86.5%, and 98.9% C.I. contours for a 2D Gaussian distribution. The dashed lines in the corresponding 1D histograms instead show the 68% C.I. of the marginalized distribution.

Appendix B

Full Hyperparameter Posteriors

In this Appendix, we report the full corner plots of the hyperparameters in both the vanilla and the redshift-dependent scenarios, shown in Figure 23 and in Figure 24, respectively. For both scenarios, we used all the selected events (286 for the

vanilla case and 80 events for the redshift-dependent), and use 10^6 injections for the selection effects computation.

In Figure 25, we report the full corner plot with a $\rho_{\text{th}} = 12$, which aligns with the criteria used for the redshift-dependent scenario. In the same figure, we also report the $\rho_{\text{th}} = 10$ results for an easier comparison.

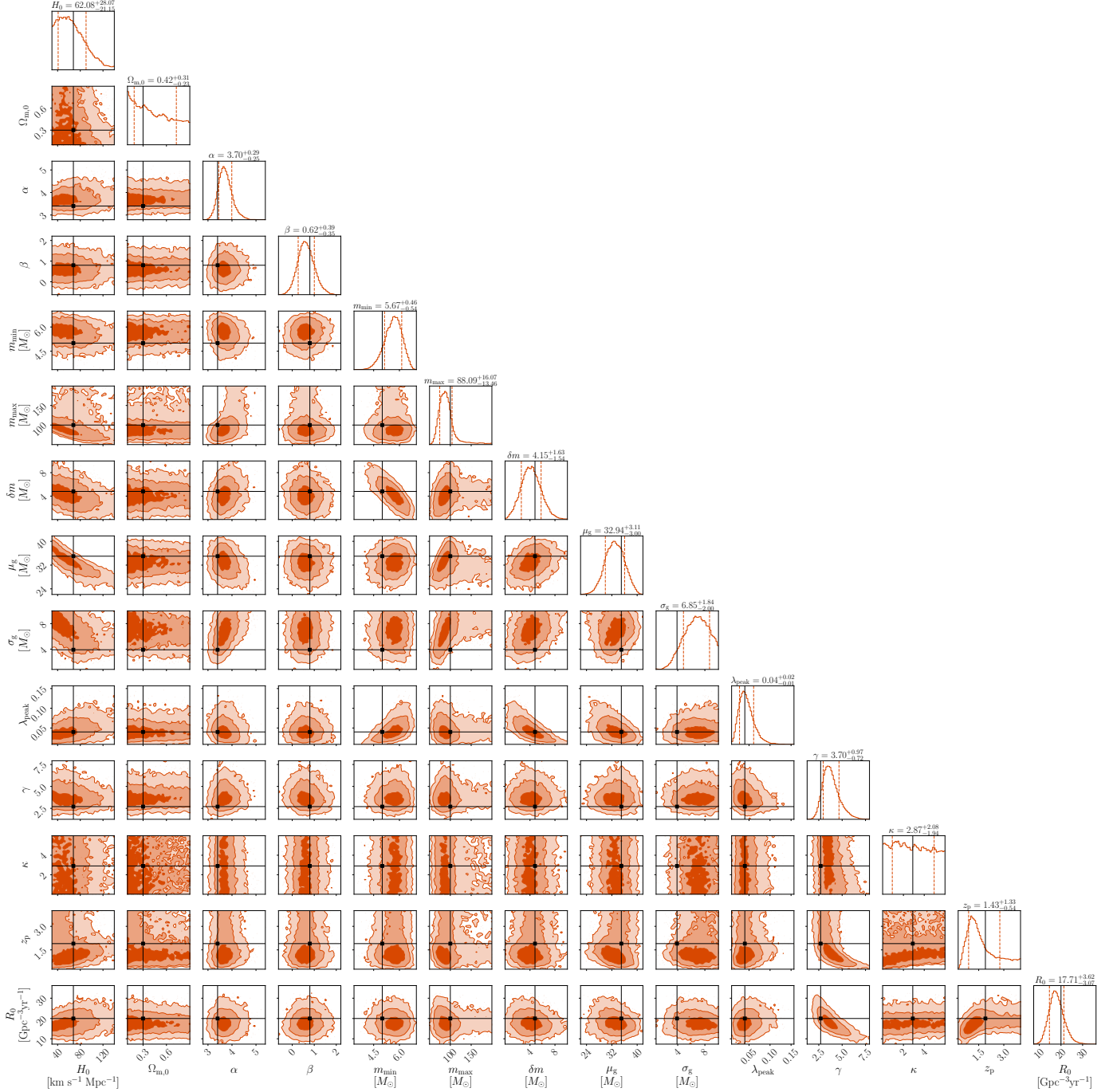


Figure 23. Posterior samples for the vanilla scenario, obtained by analyzing all the 286 events using 10^6 injections to estimate the detection probability p_{det} . The three different colored regions of the 2D posterior represent, from the darkest to the lightest shade, respectively, the 39.3%, 86.5%, and 98.9% C.I. contours for a 2D Gaussian distribution. The dashed lines in the corresponding 1D histograms instead show the 68% C.I. of the marginalized distribution.

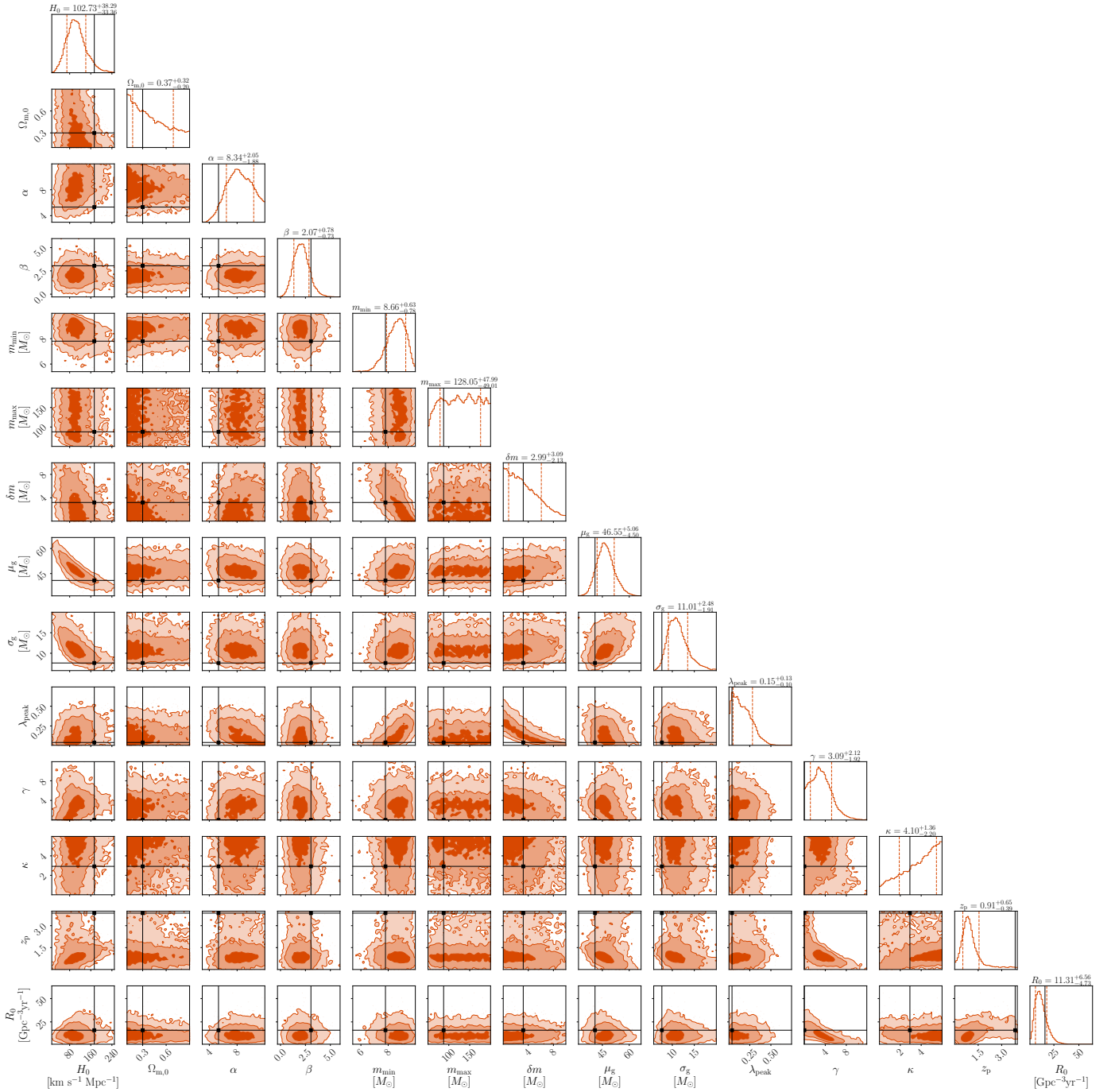


Figure 24. Posterior samples for the redshift-dependent scenario, obtained by analyzing all the 80 events using 10^5 injections to estimate the detection probability p_{det} . The three different colored regions of the 2D posterior represent, from the darkest to the lightest shade, respectively, the 39.3%, 86.5%, and 98.9% C.I. contours for a 2D Gaussian distribution. The dashed lines in the corresponding 1D histograms instead show the 68% C.I. of the marginalized distribution.

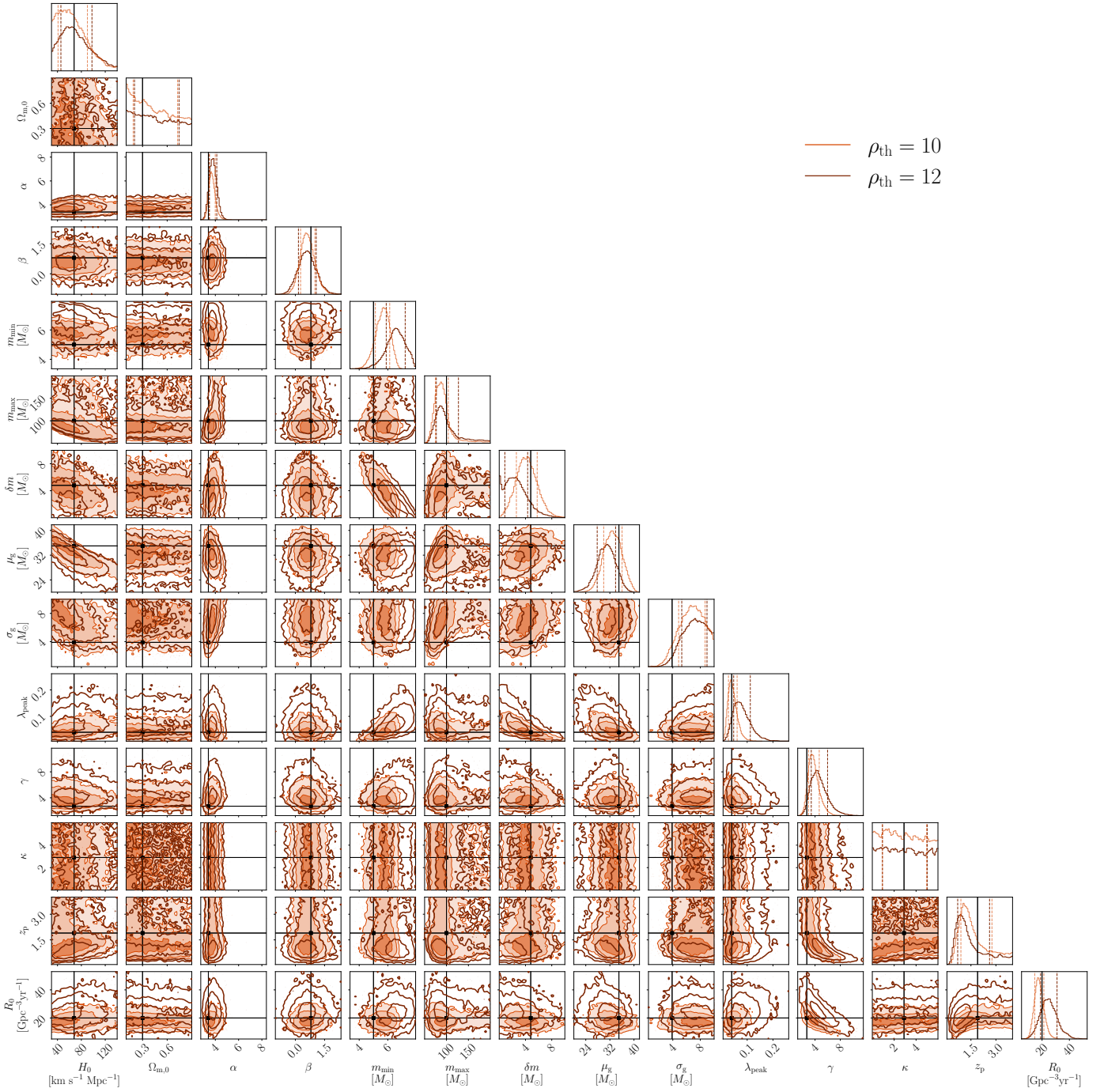


Figure 25. Posterior samples for the vanilla scenario, obtained by imposing a threshold ρ_{th} of 12 as in the redshift-dependent scenario, compared to the $\rho_{\text{th}} = 10$ as in Figure 23. In the higher threshold case, we use 10^5 injections to estimate the detection probability p_{det} . The total amount of events amounts to 164 in this case. The three different colored regions (full for $\rho_{\text{th}} = 10$ and contours only for $\rho_{\text{th}} = 12$) of the 2D posterior represent, from the darkest to the lightest shade, respectively, the 39.3%, 86.5%, and 98.9% C.I. contours for a 2D Gaussian distribution. The dashed lines in the corresponding 1D histograms instead show the 68% C.I. of the marginalized distribution.

Appendix C

Additional Analyses of the Vanilla Model

We report the full plots for the statistical tests done for the vanilla case. We refer to Section 5 for the detailed description of the tests. In summary:

1. In Figure 26, we show that changing the number of injections to evaluate selection effects between 10^4 , 10^5 , and 10^6 gives different results for all the hyperparameters.
2. In Figure 27, we show that changing the realization of the injection sample while keeping the number of injections fixed at 10^5 does not have any effect on the results.
3. In Figure 28, we show the impact of choosing a different number of events in the analysis. Apart from H_0 and

$\Omega_{m,0}$, which are slightly affected, none of the other hyperparameters show major discrepancies if the number of events is larger than 180. We notice that with 150 events only we have some deviations from the general trend.

4. In Figure 29, we show that changing the set of events with a total number fixed to 50 does not result in major discrepancies. This confirms that we do not have any particular outlier in our data set.
5. In Figure 30, which shows a companion plot to Figure 8, we did the analysis using a larger prior on both the H_0 and the σ_g parameters to see if this would have reduced the railing of H_0 on the left and the one of σ_g on the right. We see that we are getting less support for σ_g at higher values, but still the strong correlation with H_0 allows for low values for the Hubble constant.

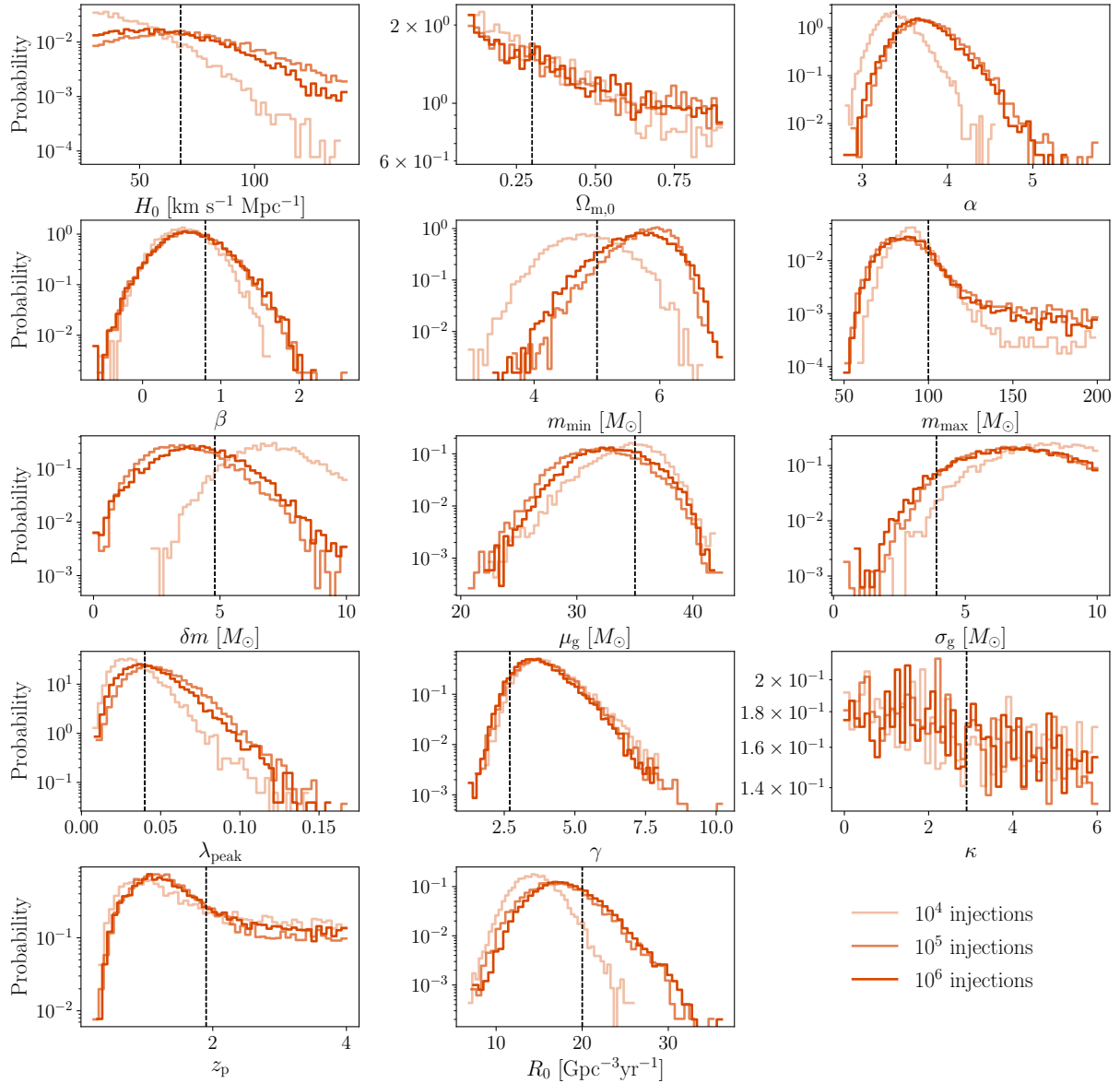


Figure 26. Vanilla case test on the dependence on the number of injections.

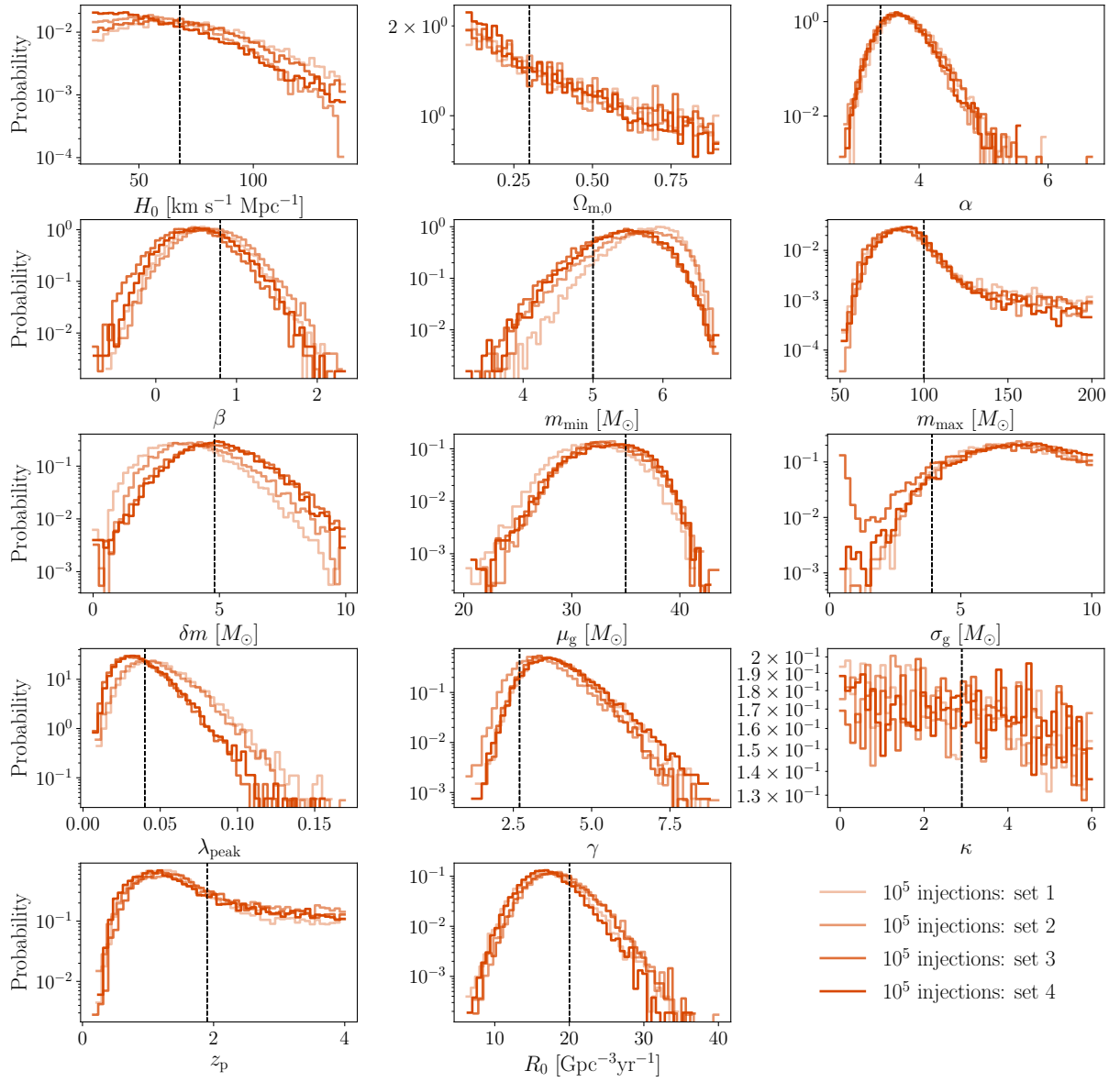
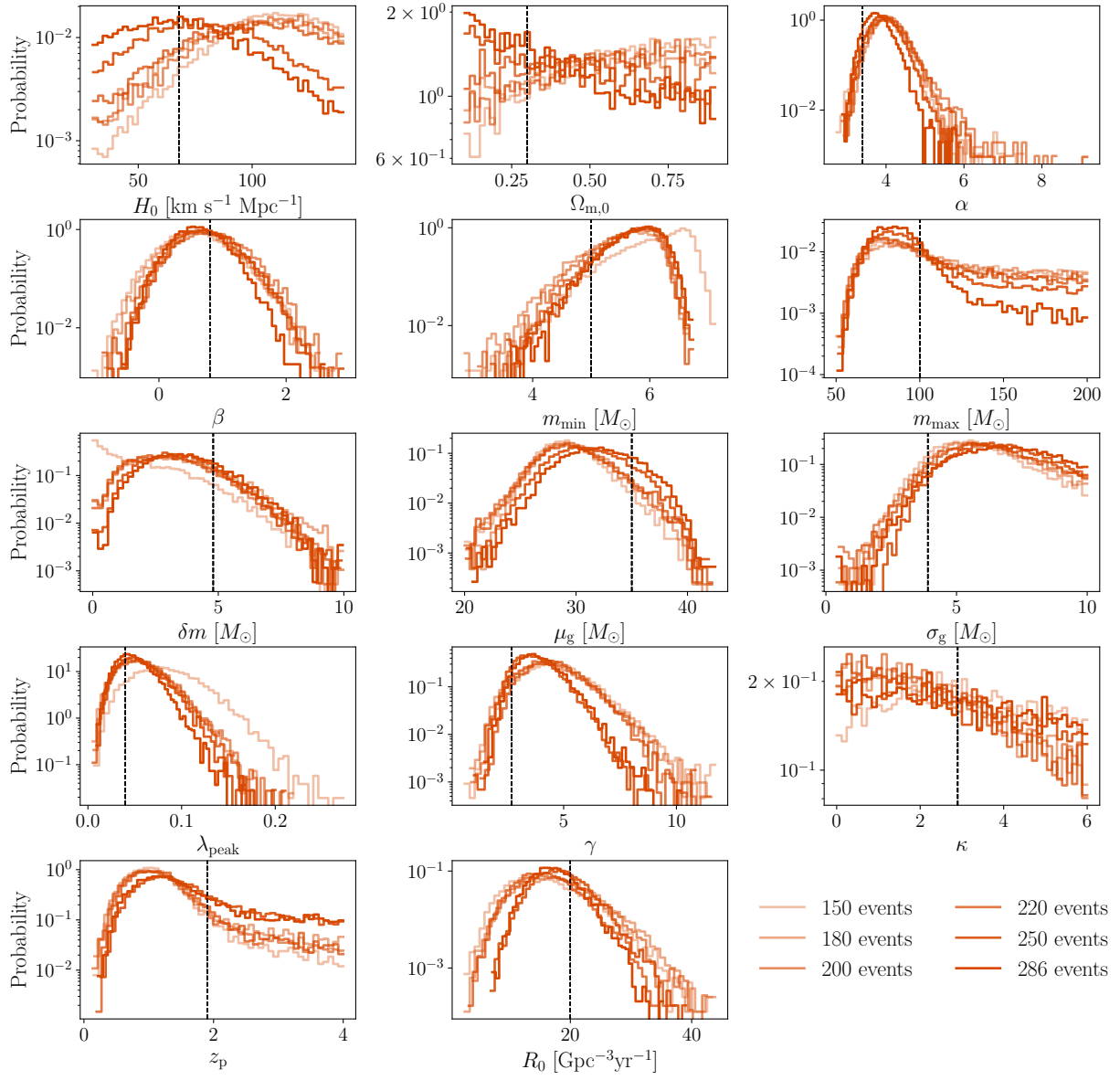


Figure 27. Vanilla case test on the dependence on the sample of injections, where the number is fixed to 10^5 .

**Figure 28.** Vanilla case test on the dependence on the number of events.

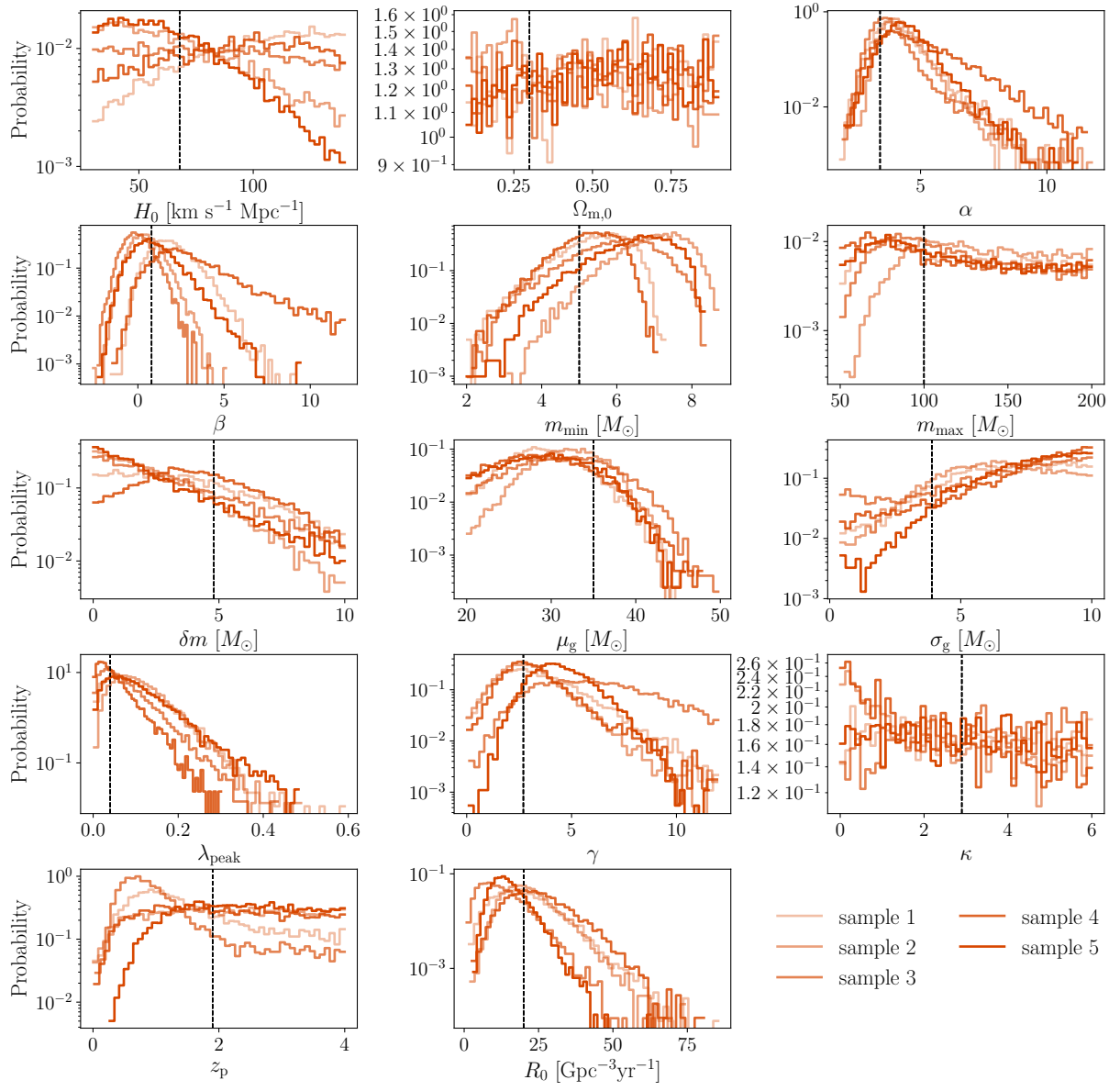


Figure 29. Vanilla case test on the dependence on the sample of events, keeping the number of events fixed to 50 in each sample.

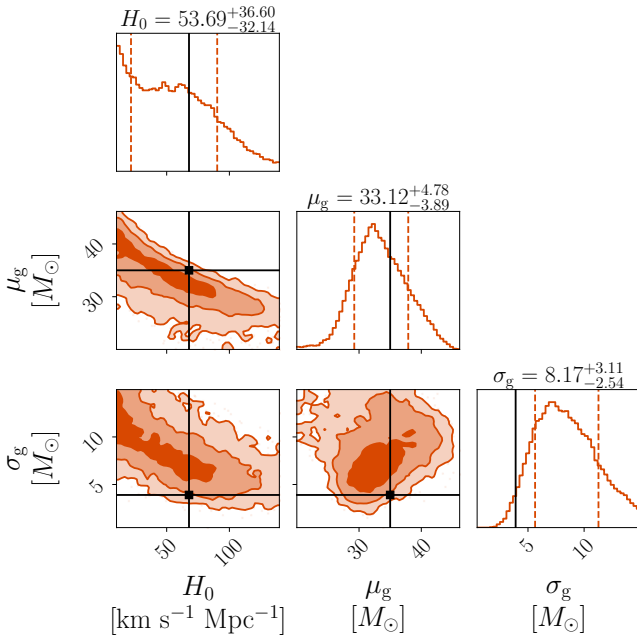


Figure 30. Vanilla case test with larger prior on H_0 , μ_g , and σ_g parameters. The injected values are represented by the solid black lines. The three different colored regions of the 2D posterior represent, from the darkest to the lightest shade, respectively, the 39.3%, 86.5%, and 98.9% C.I. contours for a 2D Gaussian distribution. The dashed lines in the corresponding 1D histograms instead show the 68% C.I. of the marginalized distribution.

ORCID iDs

Aman Agarwal <https://orcid.org/0000-0002-8685-5477>
 Ulyana Dupletsa <https://orcid.org/0000-0003-2766-247X>
 Konstantin Leyde <https://orcid.org/0000-0001-7661-2810>
 Suvodip Mukherjee <https://orcid.org/0000-0002-3373-5236>
 Benot Revenu <https://orcid.org/0000-0002-7629-4805>
 Antonio Enea Romano <https://orcid.org/0000-0002-0314-8698>
 Mohit Raj Sah <https://orcid.org/0009-0005-9881-1788>
 Sergio Vallejo-Peña <https://orcid.org/0000-0002-6827-9509>
 Adrian Avendano <https://orcid.org/0000-0002-4003-7233>
 Gergely Dalya <https://orcid.org/0000-0003-3258-5763>
 Miguel Cifuentes Espitia <https://orcid.org/0009-0006-1111-3213>
 Christos Karathanasis <https://orcid.org/0000-0002-0642-5507>
 Santiago Moreno- Gonzalez <https://orcid.org/0009-0009-8350-3247>
 Lucas Quiceno <https://orcid.org/0009-0009-2287-2102>
 Federico Stachurski <https://orcid.org/0000-0002-8658-5753>
 Juan Garcia-Bellido <https://orcid.org/0000-0002-9370-8360>
 Rachel Gray <https://orcid.org/0000-0002-5556-9873>
 Nicola Tamanini <https://orcid.org/0000-0001-8760-5421>
 Cezary Turski <https://orcid.org/0000-0001-7122-6240>

References

Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2016, *PhRvL*, **116**, 061102
 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, *Natur*, **551**, 85

Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2018, *LRR*, **21**, 3
 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**, 011048
 Abbott, R., Abe, H., Acernese, F., et al. 2023b, *ApJ*, **949**, 76
 Abdalla, E., Abellán, G. F., Aboubrahim, A., et al. 2022, *JHEAp*, **34**, 49
 Acernese, F., Agathos, M., Agatsuma, K., et al. 2015, *CQGra*, **32**, 024001
 Afroz, S., & Mukherjee, S. 2024, *MNRAS*, **534**, 1283
 Ashton, G., Hübner, M., Lasky, P. D., et al. 2019, *ApJS*, **241**, 27
 Aso, Y., Michimura, Y., Somiya, K., et al. 2013, *PhRvD*, **88**, 043007
 Bera, S., Rana, D., More, S., & Bose, S. 2020, *ApJ*, **902**, 79
 Bom, C. R., & Palmese, A. 2024, *PhRvD*, **110**, 083005
 Callister, T. A. 2024, arXiv:2410.19145
 Chen, H.-Y., Haster, C.-J., Vitale, S., Farr, W. M., & Isi, M. 2022, *MNRAS*, **513**, 2152
 Clarke, C. J., & Bromm, V. 2003, *MNRAS*, **343**, 1224
 Dehghani, A., Kim, J. L., Hosseini, D. S., et al. 2025, *JCAP*, **2025**, 056
 DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, arXiv:1611.00036
 Diaz, C. C., & Mukherjee, S. 2022, *MNRAS*, **511**, 2782
 Ding, X., Treu, T., Shajib, A. J., et al. 2018, arXiv:1801.01506
 Dopcke, G., Glover, S. C., Clark, P. C., & Klessen, R. S. 2013, *ApJ*, **766**, 103
 Dore, O., Werner, M. W., Ashby, M. L. N., et al. 2018, arXiv:1805.05489
 Ezquiaga, J. M., & Holz, D. E. 2022, *PhRvL*, **129**, 061102
 Farah, A. M., Callister, T. A., Ezquiaga, J. M., Zevin, M., & Holz, D. E. 2025, *ApJ*, **978**, 153
 Farr, W. M., Fishbach, M., Ye, J., & Holz, D. 2019, *ApJL*, **883**, L42
 Ferri, J., Tashiro, I. L., Abramo, L. R., et al. 2025, *JCAP*, **2025**, 008
 Finke, A., Foffa, S., Iacovelli, F., Maggiore, M., & Mancarella, M. 2021, *JCAP*, **2021**, 026
 Fishbach, M., Gray, R., Magaña Hernandez, I., et al. 2019, *ApJL*, **871**, L13
 Gayathri, V., Healy, J., Lange, J., et al. 2021, *ApJL*, **908**, L34
 Gray, R., Hernandez, I. M., Qi, H., et al. 2020, *PhRvD*, **101**, 122001
 Gray, R., Beirnaert, F., Karathanasis, C., et al. 2023, *JCAP*, **2023**, 023
 Hannam, M., Schmidt, P., Bohé, A., et al. 2014, *PhRvL*, **113**, 151101
 Hanselman, A. G., Vijaykumar, A., Fishbach, M., & Holz, D. E. 2025, *ApJ*, **979**, 9
 Husa, S., Khan, S., Hannam, M., et al. 2016, *PhRvD*, **93**, 044006
 Karathanasis, C., Mukherjee, S., & Mastrogianni, S. 2023a, *MNRAS*, **523**, 4539
 Karathanasis, C., Revenu, B., Mukherjee, S., & Stachurski, F. 2023b, *A&A*, **677**, A124
 Kunert, N., Gair, J., Pang, P. T. H., & Dietrich, T. 2024, *PhRvD*, **110**, 043520
 Leyde, K., Green, S. R., Toubiana, A., & Gair, J. 2024, *PhRvD*, **109**, 064056
 Leyde, K., Mastrogianni, S., Steer, D. A., Chassande-Mottin, E., & Karathanasis, C. 2022, *JCAP*, **2022**, 012
 Li, J., Liu, C., Zhang, Z.-Y., et al. 2023, *Natur*, **613**, 460
 LIGO Scientific Collaboration, Aasi, J., Abbott, B. P., et al. 2015, *CQGra*, **32**, 074001
 Loredo, T. J. 2004, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 24th Int. Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 735, ed. R. Fischer, R. Preuss, & U. V. Toussaint (Melville, NY: AIP), 195
 LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv:0912.0201
 Madau, P., & Dickinson, M. 2014, *ARA&A*, **52**, 415
 Magaña Hernandez, I., & Ray, A. 2020, arXiv:2404.02522
 Mali, U., & Essick, R. 2025, *ApJ*, **980**, 85
 Malmquist, K. G. 1922, *MeLuF*, **100**, 1
 Mandel, I., Farr, W. M., & Gair, J. R. 2019, *MNRAS*, **486**, 1086
 Mastrogianni, S., Leyde, K., Karathanasis, C., et al. 2021, *PhRvD*, **104**, 062009
 Mastrogianni, S., Pierra, G., Perriès, S., et al. 2024, *A&A*, **682**, A167
 Mokiem, M., de Koter, A., Vink, J., et al. 2007, *A&A*, **473**, 603
 Mukherjee, S. 2022, *MNRAS*, **515**, 5495
 Mukherjee, S., Ghosh, A., Graham, M. J., et al. 2020b, arXiv:2009.14199
 Mukherjee, S., Krolewski, A., Wandelt, B. D., & Silk, J. 2024, *ApJ*, **975**, 189
 Mukherjee, S., Lavaux, G., Bouchet, F. R., et al. 2021a, *A&A*, **646**, A65
 Mukherjee, S., & Wandelt, B. D. 2018, arXiv:1808.06615
 Mukherjee, S., Wandelt, B. D., Nissanke, S. M., & Silvestri, A. 2021b, *PhRvD*, **103**, 043520
 Mukherjee, S., Wandelt, B. D., & Silk, J. 2020a, *MNRAS*, **494**, 1956
 Müller, M., Mukherjee, S., & Ryan, G. 2024, *ApJ*, **977**, 45
 Nimmonkar, H., & Mukherjee, S. 2023, *MNRAS*, **527**, 2152
 Palmese, A., Bom, C. R., Mucesh, S., & Hartley, W. G. 2021, *ApJ*, **943**, 56
 Pérez-Fernández, A., Medina-Varela, L., Ruggeri, R., et al. 2025, *JCAP*, **2025**, 144
 Perna, G., Mastrogianni, S., & Ricciardone, A. 2025, *A&A*, **698**, A128
 Pierra, G., Mastrogianni, S., Perriès, S., & Mapelli, M. 2024, *Phys. Rev. D*, **109**, 083504

- Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, [A&A](#), **594**, A12
- Refregier, A., Amara, A., Kitching, T. D., et al. 2010, arXiv:1001.0061
- Salvarese, A., & Chen, H.-Y. 2024, [ApJL](#), **974**, L16
- Schutz, B. F. 1986, [Natur](#), **323**, 310
- Soares-Santos, M., et al. 2019, [ApJL](#), **876**, L7
- Speagle, J. S. 2020, [MNRAS](#), **493**, 3132
- Turski, C., Bilicki, M., Dállya, G., Gray, R., & Ghosh, A. 2023, [MNRAS](#), **526**, 6224
- van Loon, J. T. 2006, in ASP Conf. Ser. 353, *Stellar Evolution at Low Metallicity: Mass Loss, Explosions, Cosmology*, ed. H. J. G. L. M. Lamers et al. (San Francisco, CA: ASP), 211
- Verde, L., Treu, T., & Riess, A. 2019, [NatAs](#), **3**, 891
- Vink, J. S., de Koter, A., & Lamers, H. 2001, [A&A](#), **369**, 574
- Vitale, S., Gerosa, D., Farr, W. M., & Taylor, S. R. 2022, in *Handbook of Gravitational Wave Astronomy*, ed. C. Bambi, S. Katsanevas, & K. D. Kokkotas (Berlin: Springer), 45