DRAFT VERSION APRIL 26, 2024
Typeset using LaTeX **twocolumn** style in AASTeX63

A dark standard siren measurement of the Hubble constant following LIGO/Virgo/KAGRA O4a

C. R. BOM,[1,2] V. ALFRADIQUE,[1] A. PALMESE,[3] G. TEIXEIRA,[1] L. SANTANA-SILVA,[1] A. SANTOS,[1] AND P. DARC[1]

[1]*Centro Brasileiro de Pesquisas Físicas, Rua Dr. Xavier Sigaud 150, 22290-180 Rio de Janeiro, RJ, Brazil*
[2]*Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Rodovia Márcio Covas, lote J2, quadra J - Itaguaí (Brazil)*
[3]*McWilliams Center for Cosmology, Department of Physics Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, 15213, PA, USA*

ABSTRACT

We present a new constraint on the Hubble constant (H_0) from the standard dark siren method using a sample of 5 well-covered gravitational wave (GW) alerts reported during the first part of the fourth LIGO/Virgo/KAGRA observing runs in combination with standard dark sirens from the first three runs. Our methodology relies on the galaxy catalog method alone. We use the full probability density estimation of photometric redshifts derived by a deep learning method using the DESI Legacy Survey and DELVE galaxy catalogs. We add the constraints from the binary black hole mergers candidates S231226av, S231206cc, S230919bj, S230627c, and S230922g to the sample of standard dark sirens analyzed in Alfradique et al. (2024). We combine the H_0 posterior for 5 new standard sirens with other 10 previous events (3 with updated posteriors), finding $H_0 = 69.9^{+13.3}_{-12.0}$ km s^{-1} Mpc^{-1} (68% Highest Density Interval) with the catalog method alone. This result represents an improvement of $\sim 23\%$ comparing the new 15 dark siren constrain with the previous 10 dark siren constraint, and a reduction in uncertainty of $\sim 40\%$ from the combination of 15 dark and bright sirens compared with the GW170817 bright siren alone. The combination of dark and bright siren GW170817 with recent jet constraints yields H_0 of $68.0^{+4.3}_{-3.8}$ km s^{-1} Mpc^{-1}, a $\sim 6\%$ precision from Standard Sirens, reducing the previous constraint uncertainty by $\sim 10\%$.

Keywords: catalogs — cosmology: observations — gravitational waves — surveys

1. INTRODUCTION

The current $4-6\sigma$ tension in the Hubble constant (Riess et al. 2019; Planck Collaboration et al. 2018; Freedman et al. 2019; Riess et al. 2021) arises from the significant discrepancy between different cosmological probes, in particular from the Cosmic Microwave Background (CMB, Planck Collaboration et al. 2018), and those using Supernovae (SN) and Cepheids for the local distance ladder (Riess et al. 2021). New independent measurements of the Hubble constant (H_0) have the potential to shed light on this discrepancy (e.g., Verde et al. 2019; Dainotti et al. 2021; Abdalla et al. 2022), and, depending on their precision, could arbitrate the tension. Among novel probes, the standard sirens (see Schutz 1986) methodology employs gravitational wave occurrences to obtain luminosity distances. This information is used to infer cosmological parameters, most notably the Hubble constant H_0, upon integration with redshift data derived from host galaxies. This emergent probe could play an important role as it is independent of the cosmic distance ladder (Chen et al. 2018; Gray et al. 2020; Bom & Palmese 2023).

Standard sirens are categorized as "bright" in instances where an electromagnetic counterpart is definitively identified alongside a singular host galaxy, and as "dark" or "statistical" in the absence of such counterparts. Bright sirens, exemplified by GW170817 (Abbott et al. 2017), yield measurements of high precision. However, the requirement of host identification poses a series of challenges due to the wide search volume and the cadence requirements (Bom et al. 2024; Andreoni et al. 2022). Also, kilonovae have been the only widely confirmed electromagnetic sources detectable from gravitational waves, although black hole mergers are proposed to produce electromagnetic counterparts in certain circumstances with a few identified candidates (Graham et al. 2023). Therefore, the dark siren method can be applied to a larger number of events, including GW170817 (Fishbach et al. 2019), the binary black hole mergers GW170814 (Soares-Santos et al. 2019) and GW190814 (Palmese et al. 2020), and several events from the first three LIGO/Virgo observing runs (Abbott et al. 2023a). Dark standard sirens assuming the catalog method (Gair et al. 2022) rely on the position and redshift of potential host galaxies, leading to

Corresponding author: Clecio R. Bom
debom@cbpf.br

less precise results than bright sirens on a single-event basis. However, combining dark and bright sirens can enhance constraints on cosmological parameters, leveraging the abundance of events without counterparts.

The fourth LIGO/Virgo/KAGRA (LVK) Observing run (O4) began on May 24, 2023, and is scheduled to span 20 months, including two months allocated for commissioning breaks dedicated to maintenance. The initial segment of O4 referred to as O4a, finished on January 16, 2024. During O4a there have been 82 GW alerts reported, of which 80 were classified as binary black hole candidates. It is important to note that the Virgo detector was not operational during the O4a run, therefore the typical sky localizations were worse than would have been expected if it had joined the run. Meanwhile, KAGRA has participated in the run for a limited duration, albeit at a notably reduced binary neutron star (BNS) inspiral range compared to the LIGO detectors. In this work, we use a set of well-localized and confident candidate events from O4a covered by public photometric catalogs and imaging data, mainly by the Dark Energy Spectroscopic Instrument (DESI) Legacy Survey (LS; Zhou et al. 2020; Dey et al. 2019) and DECam Local Volume Exploration Survey (DELVE; Drlica-Wagner et al. 2021, 2022) , derived the photometric redshifts using full probability density estimation of photo-z using the same approach from (Alfradique et al. 2024) to adding 5 new sirens expand the sample of dark sirens from (Palmese et al. 2020; Palmese et al. 2023; Alfradique et al. 2024) to 15 sirens. The dark siren catalog method used in this work to constrain H_0 is described in the aforementioned papers, we also refer to the Appendix B for a description of our methodology.

2. DATA

2.1. *The LIGO/Virgo GW data*

In this study, we expand the analysis upon the previous works with 8-events (Palmese et al. 2023, thereafter P23) and 10 events (Alfradique et al. 2024, thereafter A24) from LVK run O1-O3, using the catalog method. We employ comparable selection criteria from prior studies (see P23 and A24) and, as such, investigate five new sirens identified in O4 with $> 70\%$ of their probability coverage falling within the Legacy or DELVE surveys, and with luminosity distances less than $d_L < 1500\,\mathrm{Mpc}$. We restricted in distance based on the previous works, as the catalogs turn more incomplete and also the dependency on Ω_m becomes more relevant. The positional data for these events comes from the maps publicly provided by the LVK collaboration, we present the 90% Credible Interval (CI) sky region of all events used in this work in Fig. 1. These maps right ascension (RA), declination (dec), and distance probability, are represented using HEALPIX pixelation. Within this framework, the probability distribution along each line of sight from the maps is assumed to follow a Gaussian distribution. We draw attention to the fact that the criteria adopted here are chosen with the intention of selecting those superevents with greater constraint capability, but there is no impediment for others to be added to the sample, as long as the selection function (see its definition in Appendix B), defined in the H_0 posterior, correctly describes the cuts adopted.

The five novel sirens candidates events we selected are S231226av (Ligo Scientific Collaboration et al. 2023e), S231206cc (LIGO Scientific Collaboration & Virgo Collaboration 2023a), S230922g (LIGO Scientific Collaboration & Virgo Collaboration 2023d), S230919bj (LIGO Scientific Collaboration & Virgo Collaboration 2023c) and S230627c (LIGO Scientific Collaboration & Virgo Collaboration 2023b). They are all classified as Binary Black Hole mergers (BBH) with probability > 99%, except for S230627c, which has a classification between neutron star-black hole (49%), BBH (48%) with (3%) of being noise. In particular, S231226av is among the lowest False-Alarm-Rate (FAR) events, while S230919bj and S230627c are in the top 20% percentile of low FAR in O4. The novel events 90% volume are comparable to the previous sirens studied with 90% volume $\sim 10^{-3}$ Gpc^3. We use the latest public skymaps (see Table 1 for details) produced by the python code Bilby (Ashton et al. 2019), that are available on the GraceDB event page [1]. We also use the last skymaps from the GW Transient Catalog (GWTC [2], LIGO Scientific Collaboration et al. 2021) for three events from P23 which are updated on this work namely GW191204_171526, GW200129_065448 and GW200311_115853.

2.2. *Optical Survey Data*

We made use of publicly available catalogs from the DESI Legacy Survey (Dey et al. 2019) and DELVE survey (Drlica-Wagner et al. 2021, 2022). The combined footprint is presented in Fig. 1. We use this data to obtain precise photometric redshifts using the same technique and Deep Learning method described in A24. We use Mixture Density Network (MDN) to derive the full probability density functions for each galaxy in the survey catalog. A more comprehensive description of the model and photometric redshift quality assessment in both surveys and a comparison between the MDN method and the public data for the Legacy Survey is presented in Appendix A. Our final constraints use the Legacy Survey Probability Density Function (PDF) instead of Gaussian approximations from the Legacy Survey photo-z catalogs (Zhou et al. 2020). The Legacy Survey results slightly outperformed the DELVE for the same model. We use DELVE-based photometric redshift as a validation

[1] https://gracedb.ligo.org/

[2] https://gwosc.org/eventapi/html/GWTC/

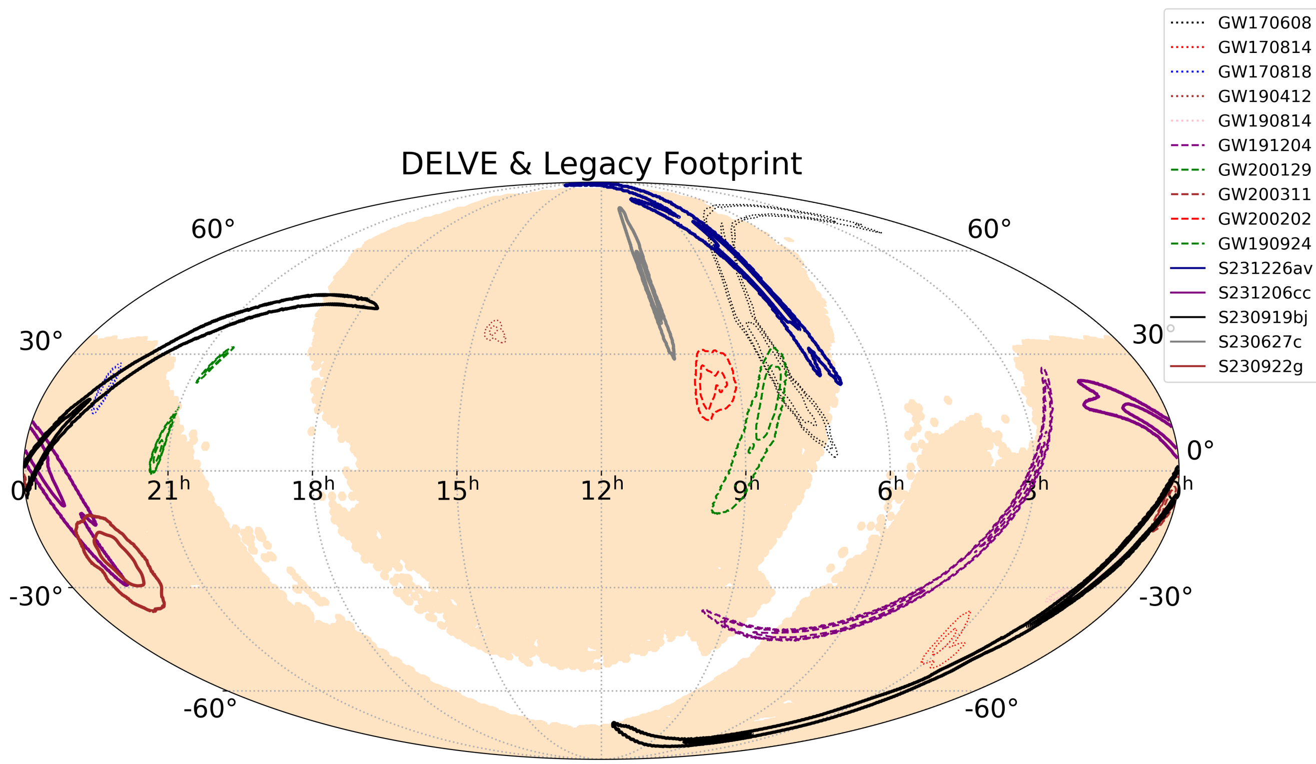

Figure 1. The LIGO/Virgo/KAGRA dark sirens analyzed in this paper. The contours depict the 90% CI localization from the sky maps. Dotted contour lines indicate events scrutinized by Palmese et al. (2023). Dashed lines represent events examined in Alfradique et al. (2024)'s study, which underwent reprocessing. Solid lines denote newly incorporated events in the analysis. The light orange shaded areas represent regions covered by the DELVE and DESI Legacy Survey catalogs.

and we found a small impact for the final constraints of $< 0.5\,\mathrm{km/s/Mpc}$.

In the GW sample processed in this work, only the S230919bj event has significant 90% region uncovered by the photo-z catalogs from both Legacy galaxy survey and DELVE. To compensate for the insufficient coverage, we adopted the same procedure from (Palmese et al. 2023). Our strategy involves distributing simulated galaxies in regions lacking data. To ensure that marginalization includes all possible host galaxies and leave our Hubble constant measurement free of an underestimated uncertainty, the injected galaxies follow our prior distribution that is given by the training sample. The photo–z distribution of these fake galaxies was sampled by the Monte Carlo technique. We assume a uniform spatial distribution, and the number density is given by the value of the Legacy Survey galaxy catalog. For the photo–z precision, we first found the relation between the photo-z error and photo-z by computing the mean and standard deviation value of the photo-z error in photo-z bins of size 0.05 from our training sample, after that, we sampled from a Gaussian function for each photo-z bins. This same procedure was adopted for the apparent magnitude distribution of these fake galaxies, since it is intrinsically associated with the redshift value. Here, we also assume that all the fake photometric redshift PDFs follow a Gaussian function; we make a Gaussian sampling around the assigned true value with the standard deviation given by their respective photo-z error. In section 3 we present the impact of these fakes galaxies and found it to be minor.

3. RESULTS AND DISCUSSION

In this section, we present the H_0 posterior produced using the dark siren methodology described in the Appendix B with the Legacy Survey photo-z's PDFs constructed using the deep learning methodology from A24 and also described in §A. We computed the H_0 posterior for five GW candidate BBH events detected in the LIGO O4a run and reanalyzed with the updated skymaps tree events from P23, as they were not yet confirmed as gravitational wave events (see §2). The 68% CI of the five O4a superevents represent about 69% to 91%, depending on the event analyzed, of the 68% CI of the prior width. The most constraining event from O4a is S230627c, because of its better localization, galaxy catalog coverage, and quality of the photo-z at the relevant redshift ranges. For the revisited gravitational wave events, despite the increase in 90% CI volume, all the three events previously studied with the `bayestar` maps in P23 present an improvement of $\sim 18\%$, 5% and 9% in 68% CI, respectively. The main cause of these improvements is the quality of our

Event	d_L [Mpc]	A [deg^2]	V [Gpc^3]	FAR
GW170608[1]	320^{+120}_{-110}	392	3×10^{-3}	<1 per 10^5 yr
GW170814[2]	540^{+130}_{-210}	62	2×10^{-3}	<1 per 10^4-10^7 yr
GW170818[1]	1060^{+420}_{-380}	39	7×10^{-3}	<1 per 10^5 yr
GW190412[3]	740^{+120}_{-130}	12	4×10^{-4}	<1 per 10^3-10^5 yr
GW190814[4]	241^{+26}_{-26}	19	3×10^{-5}	<1 per 10^4-10^7 yr
GW190924_021846[3]	564^{+145}_{-145}	348	1×10^{-2}	<1 per 10^5 yr
GW191204_171526[5]	624^{+123}_{-123}	256	8×10^{-3}	<1 per 10^5 yr
GW200129_065448[5]	929^{+179}_{-179}	31	3×10^{-3}	<1 per 10^5 yr
GW200311_115853[5]	1154^{+206}_{-206}	35	6×10^{-3}	<1 per 10^5 yr
GW200202_154313[8]	409^{+95}_{-95}	167	2×10^{-3}	<1 per 10^5 yr
S231226av[6]	1218^{+171}_{-171}	199	3×10^{-2}	<1 per 10^{42} yr
S231206cc[7]	1467^{+264}_{-264}	342	9×10^{-2}	<1 per 10^{27} yr
S230922g[8]	1491^{+443}_{-443}	324	1×10^{-1}	<1 per 10^{16} yr
S230919bj[9]	1491^{+402}_{-402}	708	2×10^{-1}	<1 per 10^2 yr
S230627c[10]	291^{+64}_{-64}	82	4×10^{-4}	<1 per 10^2 yr

Table 1. Luminosity distance, 90% CI area and volume, and False Alarm Rate (FAR) of gravitational wave events and candidates used in this analysis. We also report the reference paper or GCN that reports to the sky map used for each event. Where a range of FAR is provided, this is because multiple FAR estimates are available from multiple search algorithms. The FAR reported for the candidates is different from the confirmed events as it is estimated from the online analysis. These candidates have all recently been confirmed as gravitational wave events in LIGO Scientific Collaboration et al. (2021). References for each event: (1) Abbott et al. (2019), (2) Abbott et al. (2017b), (3) Abbott et al. (2021), (4) Abbott et al. (2020), (5) Abbott et al. (2023b), (6) Ligo Scientific Collaboration et al. (2023e), (7) Ligo Scientific Collaboration et al. (2023d), (8) Ligo Scientific Collaboration et al. (2023c) (9) Ligo Scientific Collaboration et al. (2023b), (10) Ligo Scientific Collaboration et al. (2023a)

photo-z measurement (see Fig. 4), in Appendix A we discuss these results in more detail.

The majority of the H_0 posteriors show a clear peak, except for S231226av and GW191204, which exhibit multiple peaks in different ranges of H_0 (raging from 49 to 136 km/s/Mpc). As already mentioned in previous work (Soares-Santos et al. 2019; Palmese et al. 2020; Palmese et al. 2023; Alfradique et al. 2024) these peaks are associated with the presence of overdensities regions along the line of sight (see Fig.3 in Appendix A). In the dark siren methodology, this is indicative of the host galaxy being more likely to be in these regions. Additionally, it is evident that certain events display a more prominent peak, which may stem from the presence of a prominent overdensity within an improved localization. In other words, the need for marginalization over fewer galaxies and/or the fact that the galaxies live at similar redshift, lead to a more informative H_0 posterior.

The H_0 posterior for GW191204 presents a considerable probability at $H_0 \approx 130$ km/s/Mpc, which is associated with an overdensity of galaxies at $z \approx 0.25$ (see Fig.3 in Appendix A). We also can see that this H_0 posterior begins to decrease at the high–H_0 end, due to the absence of galaxies at high redshift (the galaxy density in Fig.3 is negative at the high z end, indicating the presence of an underdense region).

The combination of the 15 dark sirens from O1-O4a is depicted in Fig.2 by the green line and detailed in table 2. The maximum *a posteriori* and the 68% CI is $H_0 = 69.9^{+13.3}_{-12.0}\,\mathrm{km/s/Mpc}$. As it can be seen, our results are consistent within 1σ with Planck (Planck Collaboration et al. 2018) and the local Cepheid-supernova distance ladder (Riess et al. 2022). This result agrees with the latest dark sirens study using the BBH population and catalog method (Abbott et al. 2023a), that found $H_0 = 67^{+13}_{-12}\,\mathrm{km/s/Mpc}$ for 47 dark sirens from the third LIGO-Virgo-KAGRA GW transient catalogue (Abbott et al. 2023b) and the `GLADE+` galaxy catalogue. A possible justification for the similar uncertainties despite the difference in number of the events used is the completeness and inhomogeneity of the `GLADE+` catalog at the redshift of interest of the studied dark sirens. This contributes to the dependence between the final H_0 constrain and the BBH population assumptions observed in (Abbott et al. 2023b), and minimizes the galaxy method contribution to the final constraint. Furthermore, we also combine our dark siren results with the bright siren GW170817 from Nicolaou et al. (2020), which considers the effect of peculiar velocity to the H_0 constrain first found by (Abbott et al. 2017a). From this analysis, we find $H_0 = 69.20^{+8.98}_{-5.85}\,\mathrm{km/s/Mpc}$. The dark siren information provides a notable reduction of 41% in the 68% CI and a 7% improvement in the relative precision. We further combine our final dark siren H_0 posterior with the bright siren analysis for GW170817 presented in Mukherjee et al. (2021), which corrects the peculiar velocity contribution in the measure of H_0 presented in Hotokezaka et al. (2019) where the superluminal motion measured by the Very Large

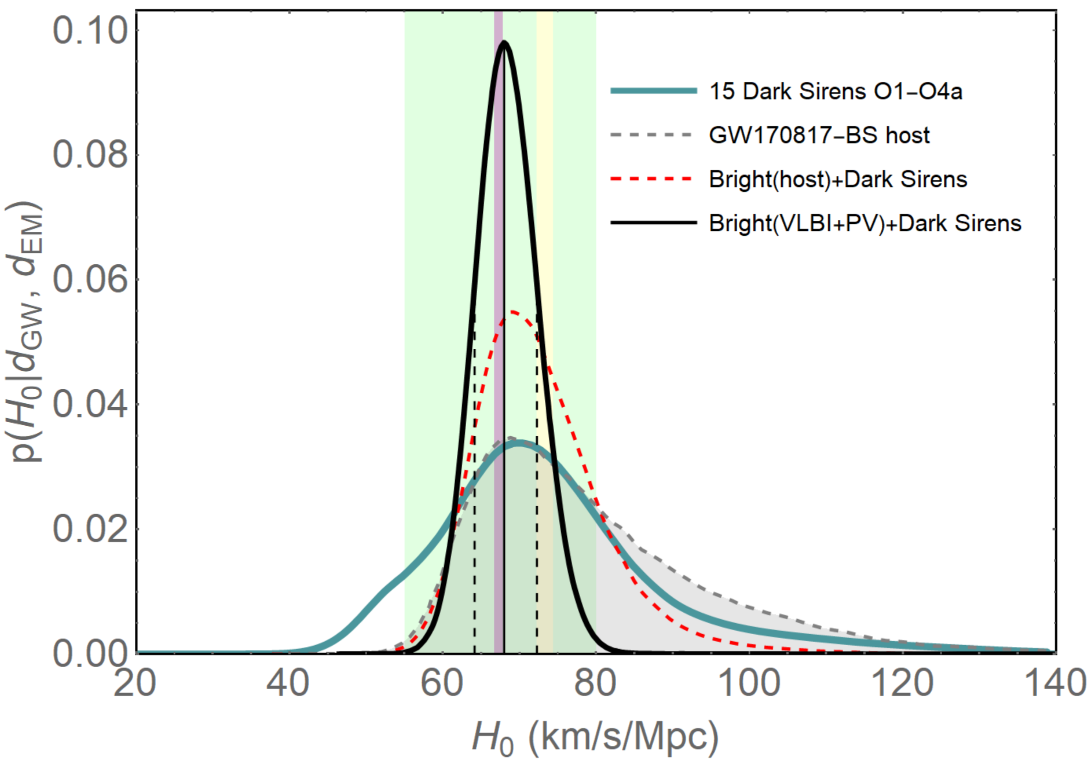

Figure 2. Hubble constant posterior distributions of the 15 dark sirens of O1-O4 observations. The dark green line shows the result from the combination of all the 15 dark sirens. The shaded gray posterior represents the GW170817 standard siren result, where only GW data is used, adapted from Nicolaou et al. (2020), which corrects the peculiar velocity to the constraint. The joint constraint from both the bright (i.e. GW170817, with the results from Nicolaou et al. (2020)) and the dark standard sirens is shown by the dotted red line. The black line is the 15 dark siren H_0 posterior (green line) combined with the results for the GW170817 bright siren found in Mukherjee et al. (2021) using the GW+VLBI data with the peculiar velocity corrections, and the vertical dashed lines show the 68% region for this posterior. For reference, we show the 1σ Planck Collaboration et al. (2018) (shaded pink), Riess et al. (2022) (R22, shaded light yellow), and the GWCT-3 (47 dark siren+ BBH population, Abbott et al. 2023a) constraints on H_0.

Baseline Interferometry (VLBI, Mooley et al. 2018) and the afterglow data is used to measure the inclination angle of the GW170817 event. The addition of an independent EM observation measure helps break the degeneracy between the distance and the inclination angle, which is one of the main contributions to the uncertainty of the gravitational wave distance measurement. We find that the dark sirens information cause a $\approx 12\%$ improvement of the 68% CI, and leads to our final constraint: $H_0 = 68.00^{+4.28}_{-3.85}\,\mathrm{km/s/Mpc}$. Furthermore, we combine our 15 dark sirens result with a more recent H_0 measurement (Palmese et al. 2024), which combines the GW170817 measurements with the electromagnetic counterpart associated with afterglow rather than the superluminal jet motion as done in (Mukherjee et al. 2021). This result presents an agreement of 1.4σ with the H_0 measurements of Planck (Planck Collaboration et al. 2018), which indicates a reduction of 0.1σ of the original result.

To guarantee the use of the most appropriate galaxies catalog in terms of coverage of the localization volumes of the gravitational wave events treated here and the quality of the estimated photo-z's, we also analyzed the measurements of H_0 for our dark sirens sample only using the DELVE catalog. This allows us to evaluate if the H_0 measurement and photo-z quality outperform our results obtained using the Legacy survey. The photometric redshifts for the DELVE galaxies were computed with the deep learning methods presented in A24 adapted for the redshift range of interest here. The results of the individual H_0 posterior distribution found with the Legacy and DELVE galaxy catalog are in agreement, the combination of the 15 dark sirens found with DELVE leads to a higher H_0 peak value (approximately 2.0 km s^{-1} Mpc^{-1}) and a lower uncertainty of H_0 ($\sigma_{H_0} \sim 0.5$ km s^{-1} Mpc^{-1}) compared to those presented in table 2. This dark siren result combined with the bright siren analysis from Mukherjee et al. (2021) gives an H_0 measurement of $68.27^{+4.20}_{-3.91}\,\mathrm{km/s/Mpc}$, also in agreement with the Legacy Survey result. Although the DELVE results present a slightly more precise measurement than those found with the Legacy Survey, the quality of the metrics of the photo-z measurements did not show the same robustness. The difference in reliability of photometric redshift measurements between DELVE and Legacy survey may be attributed to the absence of coadded data in DELVE, leading to lower signal-to-noise ratios (S/N) compared to the Legacy Survey catalog. Furthermore, Legacy offers uniform sky coverage, enabling analysis of events across both the Southern and Northern skies. Given that the precision of photometric redshifts primarily influences H_0 measurements, we selected the Legacy results as our primary outcome.

We also investigated the effect of adding galaxy fakes to the galaxy catalog for the S230919bj event. The presence of the galaxy fakes causes the peak of the posterior distribution to be slight shifted (a percent-level change) and the uncertainty is reduced by about 0.3%. Concerning the combined result of the 15 dark sirens, the presence of the galaxy fakes causes a 1% increase in precision, and the result combined with GW170817 remains unchanged, which shows the small impact that galaxy fakes have on the final conclusions of this work.

As other dark siren measurements, our H_0 measurement are valid under the Flat ΛCDM model and also presents a dependency with the background cosmology since the events go beyond $z \sim 0.1$ where changes to Ω_m and other cosmological parameters have a more significant impact on H_0 estimates. Therefore, we check if our results change with the choice of the Ω_m value within the 5σ interval found by the CMB measurement Planck Collaboration et al. (2018). There is a minor shift, relative difference $< 4\%$, in the peak of the H_0 posterior distributions, and a relative difference $< 1\%$ in the 68% CI. Therefore, the H_0 constraints presents here do not depend on the value of Ω_m, if it agrees with the Planck constraints.

The H_0 posteriors presented in this work were computed considering the full redshift PDF, computed through the deep learning algorithm described A, for each of the possible host galaxies. The application of a photo-z PDF guarantees the use of a more reliable galaxies distribution first proposed in (Palmese et al. 2020), this treatment is different from other previous (LIGO Scientific Collaboration & Virgo Collaboration 2019; Soares-Santos et al. 2019; Abbott et al. 2023a) dark sirens works that assumed a Gaussian approach for these distributions. In Appendix C, we discuss the comparison between the results of these different methodologies; the H_0 posteriors' behavior are slightly different, being more significant in closer events in which the effect of marginalization of thousands of galaxies is minimized. As already noted by Palmese et al. (2020), and also seen here, the use of the Gaussian approximation causes the distribution of galaxies dN/dz to be smoothed out, resulting in a flatter H_0 posterior. We note that the behavior of the H_0 posteriors generated by these two methodologies is slightly different, being more significant in closer events where the marginalization effect of thousands of galaxies is minimized. Despite these differences, the H_0 measurements are consistent with each other.

4. CONCLUSIONS

In this contribution, we use the data from the current best-localized and covered GW events from LVK O4a observing run to derive a dark siren measurement of the Hubble constant using the galaxy catalog method and precise photometric redshifts. We obtained $H_0 = 69.9^{+13.3}_{-12.0}$ km s^{-1} Mpc^{-1}, i.e. a $\sim 18\%$ uncertainty on H_0 from dark sirens based on the catalog method alone and a total of 15 sirens. This is, at the best of our knowledge, an unprecedented precision for the catalog method and for the dark siren approach.

We combine our result with 15 dark sirens with recent constraints over the one bright standard siren available GW170817, considering the constraints to the viewing angle from VLBI and the host galaxy peculiar velocity (Mukherjee et al. 2021) and obtained $H_0 = 68.00^{+4.28}_{-3.85}$ km s^{-1} Mpc^{-1}, a 6% measurement of H_0, reducing the previous constraint uncertainty in $\sim 10\%$. We note that the precision and uncertainty of the 15 dark sirens are similar to the GW170817 constraint without the viewing angle constraints from electromagnetic observations. It is worth noticing that current results are derived within the assumption of a Flat ΛCDM scenario.

Our current results emphasize that a combination of well-localized dark sirens and high-quality photometric redshifts can achieve a competitive H_0 constraint from gravitational waves, in particular, considering the absence of high confidence binary neutron star merger detections during O4a, the BNS merger rate can be in the lower-end of previous estimations. Therefore, the number of BBHs detections can be one order of magnitude higher than BNS. Furthermore, neutron star-black hole (NSBH) systems do not present substantial promise as multimessenger sources (Biscoveanu et al. 2023). Nonetheless, it is prudent to acknowledge that these detection rates may significantly change in the near future, and the emergence of a singular, observable BNS event with an electromagnetic counterpart could potentially offer more compelling constraints than a dozen dark standard sirens.

The current constraints from Dark Sirens by the catalog method only, achieve a precision of $\sim 18\%$ and, in combination with Bright Sirens and additional constraints in the viewing angle, can achieve $\sim 6\%$. As the number of dark sirens events increase and we get closer to the level of statistical precision required to arbitrate the Hubble tension of $\sim 2\%$ detailed studies to address potential systematics not included in this work should be carried out, especially considering different formation channels for BBH populations and catalog depth (Gray et al. 2020; Mastrogiovanni et al. 2023).

Event(s)	Method(s)	H_0 (km/s/Mpc)	σ_{H_0} (km/s/Mpc)
O1-O3a [1]	Catalog	$67.3^{+27.6}_{-17.9}$	22.5 (34%)
O1-O3 – 8 dark sirens [2]	Catalog	$79.8^{+19.1}_{-12.8}$	15.8 (20%)
O1-O3 – 47 dark sirens [3]	Catalog + BBH population	67^{+13}_{-12}	12.5 (18%)
O1-O3 – 47 dark sirens [3]	BBH population	67^{+14}_{-13}	13.5 (20%)
O1-O3 - 10 dark sirens [4]	Catalog	$76.00^{+17.6}_{-13.4}$	15.6 (20%)
GW170817 [5]	Bright (v_p corrected)	$68.80^{+17.3}_{-7.6}$	12.5(18%)
GW170817 [6]	Bright (Chandra+HST+VLA)	$75.46^{+5.34}_{-5.39}$	5.36(7%)
GW170817 [7]	Bright (v_p+VLBI)	$68.3^{+4.6}_{-4.5}$	4.6(7%)
O1-O4a – 15 dark sirens [8]	Catalog	$69.9^{+13.3}_{-12.0}$	12.6 (18%)
O1-O4a – 15 dark + 1 bright sirens [8]	Bright (host) + Catalog	$69.2^{+9.0}_{-5.8}$	7.4 (11%)
O1-O4a – 15 dark + 1 BS (EM) [8]	Bright (Chandra+HST+VLA) + Catalog	$74.3^{+5.0}_{-5.0}$	5.0 (7%)
O1-O4a – 15 dark + 1 BS (EM) [8]	Bright (v_p+VLBI) + Catalog	$68.0^{+4.3}_{-3.8}$	4.0 (6%)

Table 2. Hubble constant measurements using gravitational wave standard sirens from this work and previous works. H_0 values and uncertainties are given in $\mathrm{km\ s^{-1}\ Mpc^{-1}}$, and H_0 priors are flat, unless otherwise stated. The uncertainty from the flat prior only is derived by assuming the same H_0 maximum found in the analysis. Quoted uncertainties represent 68% HDI around the maximum of the posterior. The "$\sigma_{H_0}/\sigma_{\rm prior}$" column shows the 68% CI from the posterior divided by 68% CI of the prior width. (1) Finke et al. (2021), (2) Palmese et al. (2023), (3) Abbott et al. (2023a), (4) Alfradique et al. (2024), (5) adapted from Nicolaou et al. (2020), (6) Palmese et al. (2024), (7) Mukherjee et al. (2021), and (8) this work.

ACKNOWLEDGMENTS

CRB acknowledges the financial support from CNPq (316072/2021-4) and from FAPERJ (grants 201.456/2022 and 210.330/2022) and the FINEP contract 01.22.0505.00 (ref. 1891/22). This material is based upon work supported by NSF Grant No. 2308193. AP thanks Constantina Nicolaou for sharing the GW170817 posterior. The authors made use of Sci-Mind servers machines developed by the CBPF AI LAB team and would like to thank A. Santos, P. Russano, and M. Portes de Albuquerque for all the support in infrastructure matters.

The Legacy Surveys consist of three individual and complementary projects: the Dark Energy Camera Legacy Survey (DECaLS; NSF's OIR Lab Proposal ID 2014B-0404; PIs: David Schlegel and Arjun Dey), the Beijing-Arizona Sky Survey (BASS; NSF's OIR Lab Proposal ID 2015A-0801; PIs: Zhou Xu and Xiaohui Fan), and the Mayall z-band Legacy Survey (MzLS; NSF's OIR Lab Proposal ID 2016A-0453; PI: Arjun Dey). DECaLS, BASS and MzLS together include data obtained, respectively, at the Blanco telescope, Cerro Tololo Inter-American Observatory, The NSF's National Optical-Infrared Astronomy Research Laboratory (NSF's OIR Lab); the Bok telescope, Steward Observatory, University of Arizona; and the Mayall telescope, Kitt Peak National Observatory, NSF's OIR Lab. The Legacy Surveys project is honored to be permitted to conduct astronomical research on Iolkam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation.

The NSF's OIR Lab is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation.

This project used data obtained with the Dark Energy Camera (DECam), which was constructed by the Dark Energy Survey (DES) collaboration. Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundacao Carlos Chagas Filho de Amparo, Financiadora de Estudos e Projetos, Fundacao Carlos Chagas Filho de Amparo a Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Cientifico e Tecnologico and the Ministerio da Ciencia, Tecnologia e Inovacao, the Deutsche Forschungsgemeinschaft and the Collaborating Institutions in the Dark Energy Survey. The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energeticas, Medioambientales y Tecnologicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenossische Technische Hochschule (ETH) Zurich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciencies de l'Espai (IEEC/CSIC), the Institut de Fisica d'Altes Energies, Lawrence Berkeley National Laboratory, the Ludwig-Maximilians Universitat Munchen and

the associated Excellence Cluster Universe, the University of Michigan, the National Optical Astronomy Observatory, the University of Nottingham, the Ohio State University, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, and Texas A&M University.
BASS is a key project of the Telescope Access Program (TAP), which has been funded by the National Astronomical Observatories of China, the Chinese Academy of Sciences (the Strategic Priority Research Program "The Emergence of Cosmological Structures" Grant # XDB09000000), and the Special Fund for Astronomy from the Ministry of Finance. The BASS is also supported by the External Cooperation Program of Chinese Academy of Sciences (Grant # 114A11KYSB20160057), and Chinese National Natural Science Foundation (Grant # 11433005).
The Legacy Survey team makes use of data products from the Near-Earth Object Wide-field Infrared Survey Explorer (NEOWISE), which is a project of the Jet Propulsion Laboratory/California Institute of Technology. NEOWISE is funded by the National Aeronautics and Space Administration.
The Legacy Surveys imaging of the DESI footprint is supported by the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy under Contract No. DE-AC02-05CH1123, by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract; and by the U.S. National Science Foundation, Division of Astronomical Sciences under Contract No. AST-0950945 to NOAO.
The Photometric Redshifts for the Legacy Surveys (PRLS) catalog used in this paper was produced thanks to funding from the U.S. Department of Energy Office of Science, Office of High Energy Physics via grant DE-SC0007914.

Software: `astropy` (Astropy Collaboration et al. 2013, 2018), `BAYESTAR` (Singer & Price 2016; Singer et al. 2016; Singer et al. 2016), `TOPCAT` (Taylor 2005)

REFERENCES

Abbott, B., Abbott, R., Abbott, T., et al. 2019, Physical Review X, 9

Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2017, Phys. Rev. Lett., 119, 161101

Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2017a, Nature, 551, 85, arXiv:1710.05835

Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2017b, Phys. Rev. Lett., 119, 141101, arXiv:1709.09660

Abbott, R., Abbott, T. D., Abraham, S., et al. 2020, ApJ, 896, L44

Abbott, R., Abe, H., Acernese, F., & et al. 2023a, Astrophys. J., 949, 37

Abbott, R., Abbott, T., Abraham, S., et al. 2021, Phys. Rev. X, 11

Abbott, R., Abbott, T. D., Acernese, F., et al. 2023b, Phys. Rev. X, 13, 041039

Abdalla, E., Abellán, G. F., Aboubrahim, A., et al. 2022, JHEAP, 34, 49

Alfradique, V., Bom, C. R., Palmese, A., et al. 2024, MNRAS, 528, 3249–3259

Andreoni, I., Margutti, R., Salafia, O. S., et al. 2022, ApJS, 260, 18

Ashton, G., Hübner, M., Lasky, P. D., Talbot, C., & et al. 2019, ApJS, 241, 27

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33

Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123

Benitez, N., Dupke, R., Moles, M., et al. 2014, J-PAS: The Javalambre-Physics of the Accelerated Universe Astrophysical Survey, arXiv:1403.5237

Biscoveanu, S., Landry, P., & Vitale, S. 2023, MNRAS, 518, 5298, arXiv:2207.01568

Bishop, C. 1994, Aston University, 26

Bom, C. R. & Palmese, A. 2023, arXiv e-prints, arXiv:2307.01330, arXiv:2307.01330

Bom, C. R., Annis, J., Garcia, A., et al. 2024, ApJ, 960, 122, arXiv:2302.04878

Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, The Astrophysical Journal, 686, 1503

Chen, H.-Y., Fishbach, M., & Holz, D. E. 2018, Nature, 562, 545, arXiv:1712.06531

Dainotti, M. G., De Simone, B., Schiavone, T., et al. 2021, The Astrophysical Journal, 912, 150

Dalmasso, N., Pospisil, T., Lee, A., et al. 2020, Astronomy and Computing, 30, 100362

Dawid, A. P. 1984, J R Stat Soc Ser A Stat, 147, 278

Dey, A., Schlegel, D. J., Lang, D., et al. 2019, AJ, 157, 168, arXiv:1804.08657

Drlica-Wagner, A., Carlin, J. L., Nidever, D. L., et al. 2021, ApJS, 256, 2, arXiv:2103.07476

Drlica-Wagner, A., Ferguson, P. S., Adamów, M., et al. 2022, ApJS, 261, 38, arXiv:2203.16565

D'Isanto, A. & Polsterer, K. L. 2018, A&A, 609, A111

Finke, A., Foffa, S., Iacovelli, F., Maggiore, M., & Mancarella, M. 2021, Cosmology with LIGO/Virgo dark sirens: Hubble parameter and modified gravitational wave propagation, arXiv:2101.12660

Fishbach, M., Gray, R., Magaña Hernandez, I., et al. 2019, ApJL, 871, L13, arXiv:1807.05667

Freedman, W. L., Madore, B. F., Hatt, D., et al. 2019, ApJ, 882, 34

Gair, J. R., Ghosh, A., Gray, R., et al. 2022, arXiv e-prints, arXiv:2212.08694, arXiv:2212.08694

Graham, M. J., McKernan, B., Ford, K. E. S., et al. 2023, The Astrophysical Journal, 942, 99

Gray, R., Hernandez, I. M., Qi, H., et al. 2020, PhRvD, 101, 122001, arXiv:1908.06050

Hermans, J., Delaunoy, A., Rozet, F., et al. 2022, A Trust Crisis In Simulation-Based Inference? Your Posterior Approximations Can Be Unfaithful, arXiv:2110.06581

Hotokezaka, K., Nakar, E., Gottlieb, O., et al. 2019, Nature, 3, 940–944, arXiv:1806.10596

Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, A&A, 457, 841

Li, C., Zhang, Y., Cui, C., et al. 2022, MNRAS, 518, 513

LIGO Scientific Collaboration & Virgo Collaboration. 2019, A gravitational-wave measurement of the Hubble constant following the second observing run of Advanced LIGO and Virgo, arXiv:1908.06060

LIGO Scientific Collaboration & Virgo Collaboration. 2023a, GRB Coordinates Network, 35298, 1

LIGO Scientific Collaboration & Virgo Collaboration. 2023b, GRB Coordinates Network, 34086, 1

LIGO Scientific Collaboration & Virgo Collaboration. 2023c, GRB Coordinates Network, 34739, 1

LIGO Scientific Collaboration & Virgo Collaboration. 2023d, GRB Coordinates Network, 34757, 1

LIGO Scientific Collaboration, Virgo Collaboration, & KAGRA Collaboration. 2021, GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run, arXiv:2111.03606

Ligo Scientific Collaboration, VIRGO Collaboration, & Kagra Collaboration. 2023a, GRB Coordinates Network, 34087, 1

Ligo Scientific Collaboration, VIRGO Collaboration, & Kagra Collaboration. 2023b, GRB Coordinates Network, 34739, 1

Ligo Scientific Collaboration, VIRGO Collaboration, & Kagra Collaboration. 2023c, GRB Coordinates Network, 34757, 1

Ligo Scientific Collaboration, VIRGO Collaboration, & Kagra Collaboration. 2023d, GRB Coordinates Network, 35298, 1

Ligo Scientific Collaboration, VIRGO Collaboration, & Kagra Collaboration. 2023e, GRB Coordinates Network, 35428, 1

Lima, E., Sodré, L., Bom, C., et al. 2022, Astron. Comput., 38, 100510

Madau, Piero & Dickinson, M. 2014, Annu. Rev. Astron. Astrophys., 52, 71

Mastrogiovanni, S., Laghi, D., Gray, R., et al. 2023, Phys. Rev. D, 108, 042002

Mooley, K. P., Deller, A. T., Gottlieb, O., et al. 2018, Nature, 561, 355–359

Mucesh, S., Hartley, W. G., Palmese, A., et al. 2021, MNRAS, 502, 2770

Mukherjee, S., Lavaux, G., Bouchet, F. R., et al. 2021, A&A, 646, 11

Nicolaou, C., Lahav, O., Lemos, P., Hartley, W., & Braden, J. 2020, MNRAS, 495, 90, arXiv:1909.09609

Palmese, A., Bom, C. R., Mucesh, S., & Hartley, W. G. 2023, ApJ, 943, 56

Palmese, A., Kaur, R., Hajela, A., et al. 2024, Phys. Rev. D, 109, 063508

Palmese, A., deVicente, J., Pereira, M. E. S., et al. 2020, ApJL, 900, L33, arXiv:2006.14961

Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2018, ArXiv e-prints, arXiv:1807.06209

Polsterer, K. L., D'Isanto, A., & Gieseke, F. 2016, Uncertain Photometric Redshifts, arXiv:1608.08016

Riess, A. G., Casertano, S., Yuan, W., et al. 2021, ApJ, 908, L6

Riess, A. G., Casertano, S., Yuan, W., Macri, L. M., & Scolnic, D. 2019, ApJ, 876, 85

Riess, A. G., Yuan, W., Macri, L. M., et al. 2022, ApJL, 934, L7, arXiv:2112.04510

Schutz, B. F. 1986, Nature, 323, 310

Singer, L. P. & Price, L. R. 2016, Phys. Rev. D, 93, 024013

Singer, L. P., Chen, H.-Y., Holz, D. E., et al. 2016, ApJ, 829, L15

Singer, L. P., Chen, H.-Y., Holz, D. E., et al. 2016, ApJS, 226, 10, arXiv:1605.04242

Soares-Santos, M., Palmese, A., Hartley, W., et al. 2019, ApJL, 876, L7, arXiv:1901.01540

Taylor, M. B. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 29

Verde, L., Treu, T., & Riess, A. G. 2019, Nat. Astron., 3, 891, arXiv:1907.10625

Voelker, A. R., Kajić, I., & Eliasmith, C. 2019, in Advances in Neural Information Processing Systems, 15544

Zhou, R., Newman, J. A., Mao, Y.-Y., et al. 2020, The Clustering of DESI-like Luminous Red Galaxies Using Photometric Redshifts, arXiv:2001.06018

Zou, H., Gao, J., Zhou, X., & Kong, X. 2019, ApJS, 242, 8

APPENDIX

A. GALAXY CATALOGS AND PHOTOMETRIC REDSHIFTS

In this work, we use catalog data from both DEcam Local Volume Exploration survey (DELVE; Drlica-Wagner et al. 2021, 2022) and the Legacy Survey (henceforth LS; Dey et al. 2019) as the main data used to select the galaxies within the LIGO/Virgo localization for each event. It is worth mentioning that the Legacy Survey catalogs incorporated publicly available DECam data, including DELVE footprint for the LS DR10. The DELVE survey (Drlica-Wagner et al. 2021, 2022) observed a large fraction of the Southern sky covering an area of 21,000 deg^2, of which 17,000 deg^2 were homogeneously observed in the four broad bands (g, r, i, and z) and photometric depth up to 24.3, 23.9, 23.5, and 22.8 for g,r, i, and z, respectively for 5σ point source detections. The DELVE catalogs comprise 2.5 billion sources, in which 618 million have data in all the four band. The Legacy survey explored a significant portion of the sky ($\sim 33,500$ sq deg) in grz bands, reaching depths of 24.0, 23.4, and 22.5 in grz for 5 σ detection. The sky coverage of DELVE and Legacy surveys catalogs are shown in Fig. 1, together with the 50 and 90% CI of the GW events studied in previous works and the five new events explored in this study.

Photometric redshifts used in this work were determined using the same deep learning method presented in (Alfradique et al. 2024). The fundamental component of the Deep Learning model comprises a neural network that analyzes tabular data, using a Legendre Memory Unit (LMU, Voelker et al. 2019) with a Mixture Density Neural Network (MDN, Bishop 1994). Unlike traditional neural networks that provide single value estimations, MDNs outputs conditional probability densities through a linear combination of individual probability distributions (components), chosen to be Normal distributions in our case. This approach enables a more comprehensive characterization of predictions and errors assessment.

The neural network output is a linear combination of C Gaussian kernels ($\mathcal{N}(\mu_i,\sigma_i)$, where $\{\mu_i\}$ is the mean and $\{\sigma_i\}$ is the standard deviation) weighted by mixture coefficients $\{\alpha_i\}$. We impose, for the mixture coefficients, that $\sum_{i=1}^{C}\alpha_i = 1$ and $0 < \alpha_i < 1$. Therefore, the PDF can be written as

$$\mathrm{PDF(z)} = \sum_{\mathrm{i=1}}^{\mathrm{C}} \alpha_{\mathrm{i}}\mathcal{N}(\mu_{\mathrm{i}},\sigma_{\mathrm{i}}). \tag{A1}$$

This method was implemented to estimate photo–z's PDFs for the DELVE DR2 and LS DR10 catalogs. The architecture used for the DELVE photo–z is detailed in Alfradique et al. (2024). The LS DR10[3] photo–z where estimated using a similar architecture, with minor adjustments after a fine tunning considering the point statistics metrics such as reducing the width of the dense layers and reducing the number of components C in the mixture, from $C = 20$ for DELVE to $C = 6$ for Legacy. Input features for this process include $griz$ magnitudes and color indices (g-r, g-i, g-z, r-i, r-z, and i-z). The spectroscopic data used as training set for DELVE catalogs is the same as the one presented in Alfradique et al. (2024), which was created from the crossmatch between DELVE catalog and several spectroscopic data available in different large sky surveys. For the Legacy survey we use the same training sample from the public photometric redshift catalog (Zhou et al. 2020).

The LS public data releases do not directly provides the apparent magnitude in photometric band. Therefore, we initially used the linear fluxes (columns `FLUX_{G,R,I,Z}`) to compute the magnitudes m by employing the conversion $m = 22.5 - 2.5\log_{10}(f)$, and derived the magnitude errors also from the inverse variances of the fluxes (columns `FLUX_IVAR_{G,R,I,Z}`) [4]. To mitigate star contamination, we adopted the same approach outlined by Palmese et al. (2023), applying color cuts based on GAIA data, such as removing all known stars. Additionally, all the magnitudes were corrected for Milky Way extinction. Finally, we restrict our GW analysis to r band magnitudes to be lower than 21.

We conducted a series of selection cuts based on the photometry quality and properties to ensure the utilization of galaxies with the best possible detections. We excluded all objects with unphysical colors, retaining only those satisfy the conditions:

$$-1 > g-r,\ r-i, i-z < 4$$

For the spectroscopic sample (refer to Alfradique et al. 2024 for a detailed list of the spectroscopic catalogs also used in this work), we restricted our objects to $0.01 < z_{\mathrm{spec}} < 1.5$. After applying these cuts, our spectroscopic sample contains $\sim 2.2M$ galaxies. We selected the training sample in order to have a uniform z_{spec} distribution of 0.01 to 1.0, and comprise all objects

[3] We used the latest avaliable version of DR10, namely 10.1

[4] https://www.legacysurvey.org/dr9/description/photometry

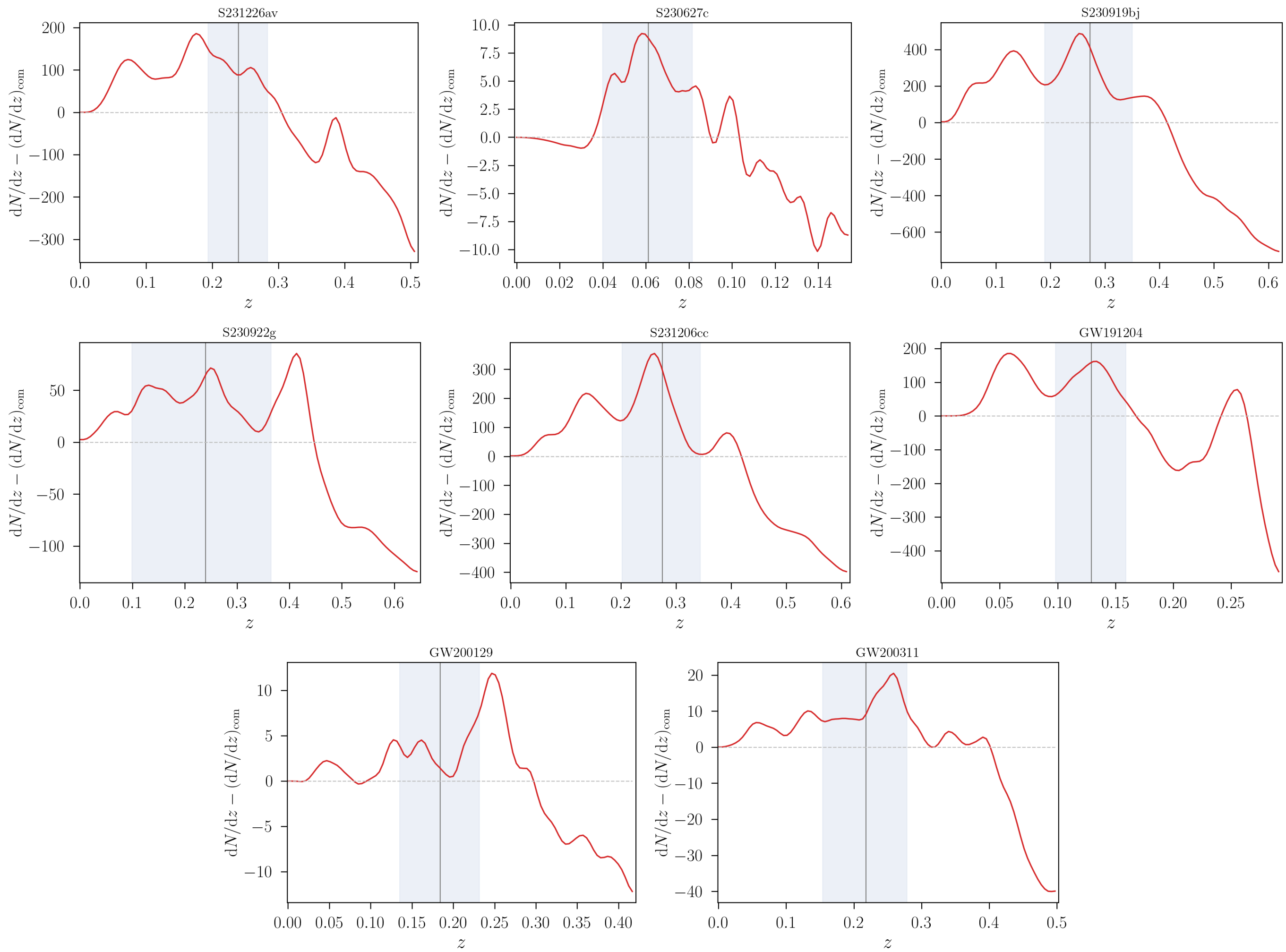

Figure 3. Redshift distribution of galaxies in the 90% CI area of the five new dark siren events from O4a (S231226av, S230627c, S230919bj, S230922g, and S231206cc), and the three superevents discussed in Palmese et al. (2023) that have now been confirmed as gravitational waves (GW1901204, GW200129 and GW200311). To highlight the presence of overdensities and underdensities along the line of sight, the redshift distribution was subtracted with a uniform number density. The grey vertical lines represent the luminosity distance of each GW event marginalized over the entire sky, assuming an H_0 of 70 km s^{-1} Mpc^{-1}, and the shaded regions are the 1σ uncertainties considering the same H_0. These regions are only showed for reference.

available with $z_{\rm spec} > 1$ (which are few in number), resulting in 550k (580k for DR9) galaxies for training the model. The same approach of uniform training was also used in Legacy Survey publicly available photo-z (Zou et al. 2019) to avoid possible systematic bias towards oversampled regions.

We generate PDFs for galaxies in the test sample to validate the model by checking probabilistic and marginal calibration. Finally, we use the trained model to generate PDFs for galaxies in our target datasets. We explore the performance of the full PDF estimations by examining both the point-estimates photometric redshifts and the calibration of their PDFs. For a given galaxy, the photo–z value is defined as its respective PDF's most probable value (peak). Fig. 3 presents the final photo-z distribution, found with the full photo-z PDF described in this appendix, in the 90% CI area of each gravitational wave event studied here. In order to highlight the overdensity regions, the photo-z distribution dN/dz was subtracted from the uniform distribution in comoving volume $(dN/dz)_{\rm com}$.

To evaluate the quality of the point estimates, we employ the following metrics:

- **Median Bias**: the bias, defined as $\Delta z = z_{\rm phot} - z_{\rm spec}$, directly measures the deviation of our estimations from the target values ($z_{\rm spec}$).

- **Scatter**: the Normalized Median Absolute Deviation $\sigma_{\rm NMAD}$, defined as

$$\sigma_{\mathrm{NMAD}} = 1.48 \times \mathrm{median}\left(\left|\frac{\Delta z - \mathrm{median}(\Delta z)}{1+z_{\mathrm{spec}}}\right|\right), \tag{A2}$$

is a standard measurement of the bias scattering (Brammer et al. 2008; Li et al. 2022; Lima et al. 2022). We aim for σ_{NMAD} to be as low as possible. The choice of σ_{NMAD} instead of σ_{68} is less sensitive to outliers.

- **Outlier fraction**: outliers are defined as objects which

$$\frac{\Delta z}{1+z_{\mathrm{spec}}} > 0.15. \tag{A3}$$

We define η as being the fraction of outliers in any sub-sample of photo–z estimations. This definition of an outlier follows the same approach as adopted in Ilbert, O. et al. (2006) and Lima et al. (2022).

All the aforementioned metrics were computed using the objects with spectroscopic correspondence inside the 90% confidence area of the events S231206cc, S230919bj, S230922g, GW191204_171526, GW190924_021846, GW200129_065458, GW200202_154313 and GW200311_115853. Additionally, with the exception of the odds constraints, the analysis refer only to the objects that satisfies the restrictions imposed to the events' photo–z catalogs.

Fig. 4 presents the median bias, scatter, and outlier fraction, respectively, as a function of photo–z bins of 0.025 width. We computed the mean value of the metrics for each bin, generating one curve per event and then averaging these curves for all the aforementioned events. The solid lines represent the mean values across all the events, while the shaded regions represent the associated standard deviation.

Considering the measurements for both DELVE and LS surveys, we obtain σ_{NMAD} and outlier fraction equal to 0.032 (0.016), 6.2(2.4)% for DELVE (Legacy), respectively. This results show that our measurements using Legacy data outperforms those obtained using DELVE data. We use the Legacy public photometric redshifts in Fig. 4 catalogs for performance comparison in the point-like metrics. The LS public photometric redshifts were computed using the Random Forest (RF) method by Zhou et al. (2020). In this method, they used r-band magnitude, *g-r*, *r-z*, *z-W1*, and *W1-W2* colors as input features. We compare our results with the public photo–z catalog from LS-DR10 as well as the public redshifts from LS-DR9 (Zou et al. 2019) utilized in Palmese et al. (2023). On the left panel of Fig. 4 the MDN method exhibits an improvement of the averaged median bias in the $0.01 < z < 0.1$ interval, compared both to the DR10 public available data and the data used in Palmese et al. (2023), which is relevant since a considerable amount of the objects in the event skymaps lies in that interval. Conversely, for higher redshifts, both curves are compatible within the scatter. Throughout the entire range of photo–z , the bias for our method is very low and does not exceed 0.007. It is also consistent with bias 0 within the scatter until z$\sim$ 0.4. However, it tends to be positive everywhere, indicating a weak cumulative overestimation effect. Additionally, in Fig.4 the scatter and outlier fraction align with publicly available redshifts values, thereby affirming the reliability of our method in generating accurate photo–z estimations.

Fig. 5 illustrates our estimations plotted against the z_{spec} values available for all the events in this analysis. Despite the feature around $z = 0.15$, we observe well-distributed predictions around the diagonal line $y = x$. The averaged median bias for this entire sample is 0.002, while the averaged σ_{NMAD} and outlier fraction are 0.014 and 0.8%, indicating the robustness of our results.

To validate the Probability Density Functions, we investigated the following metrics:

- **PIT**:

  the Probability Integral Transform (PIT, Dawid 1984) represents the Cumulative Density Function of the PDF up to the z_{spec} value for each galaxy. We can assess the quality of the PDFs by examining the PIT distribution for a representative sample of galaxies. The distribution of PIT values is expected to be uniform between 0 and 1 (Mucesh et al. 2021), indicating that the z_{spec} values can be considered as random events generated from independent PDFs. A slope in the PIT distribution indicates a bias in the estimations, while the concavity of the distribution reveals whether the PDFs are over- or under-dispersed (Polsterer et al. 2016).

- **Odds**:

  the odds (Benitez et al. 2014) measure the degree of confidence in the photo-z estimate derived from a given PDF. It is defined as the probability of the redshift lying within an interval around the photo-z value. For a given PDF, we can compute:

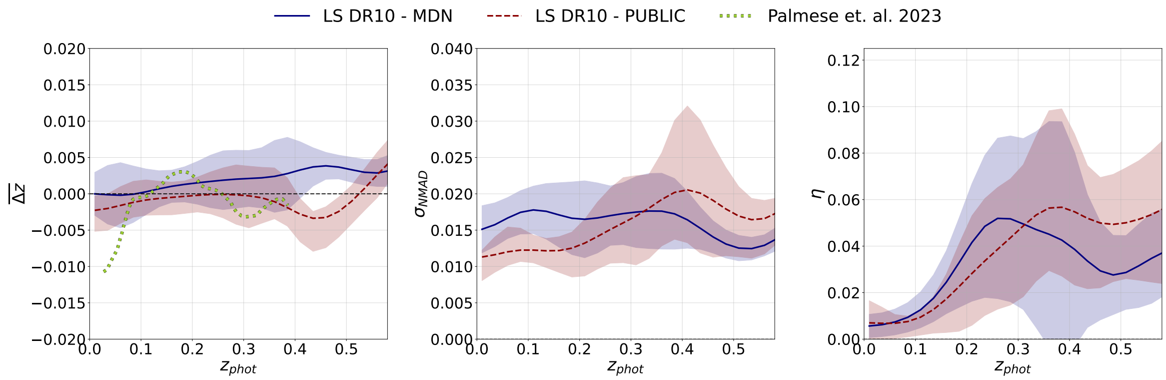

Figure 4. Mean values of the median bias (Δz), $\sigma_{\rm NMAD}$, and η as a function of 0.025 width (photo–z) bins. The solid lines depict the computed averages for all objects within each bin for each metric. The shadows represent the respective standard deviations.

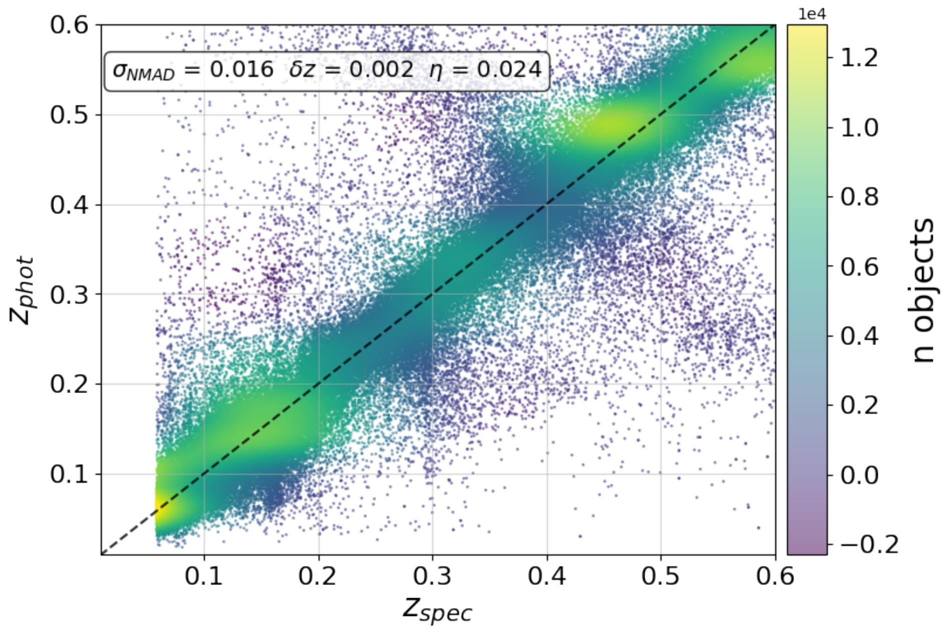

Figure 5. Photo-z's estimations versus the $z_{\rm spec}$ values. The colormap represents the density of objetcs, and on the top of the figure we displayed the point estimation metrics for the entire specz sample within the analyzed GW and superevent regions.

$$\text{Odds} = \int_{\rm z_{phot}-0.06}^{\rm z_{phot}+0.06} \text{PDF(z)dz}. \tag{A4}$$

The ideal distribution of odds for a galaxy sample well-represented by the training set should exhibit a pronounced peak near 1. However, it could also represent a distribution of under-dispersed PDFs. For this reason, we have to analyze the odds and PIT distribution simultaneously, in order to infer the quality of the PDFs.

- **Coverage diagnostic**:

the coverage diagnostic (or High posterior density, HPD, diagnostic) stands as a well-established metric commonly employed to assess the quality of credible regions generated by simulation-based inference algorithms. Additionally, it provides a means to evaluate the accuracy of the estimated redshift distribution, as outlined in Dalmasso et al. (2020). The fundamental concept involves evaluating the probability that a specified credible region within the inferred distribution contains the true value. This assessment provides insights into whether the estimated distribution is overconfident, cali-

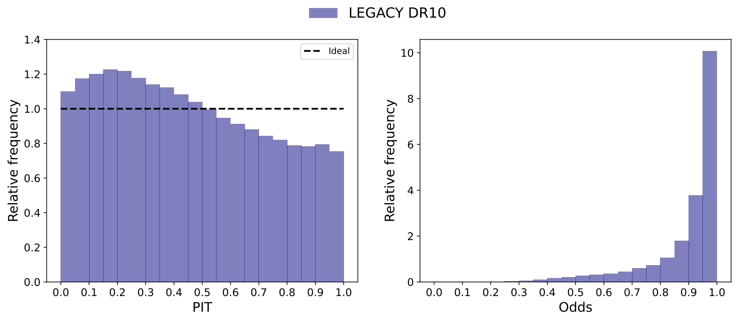

Figure 6. PIT and odds distributions for the objects with spectroscopic redshifts within the event regions. The odds distributions (right panel) remains for all objects (around 245k galaxies), meanwhile the PIT distribution (left panel) remains for the objects which obey $Odds > 0.7$ beyond the errors constrains employed to the $H0$ posterior estimations (around 130k galaxies).

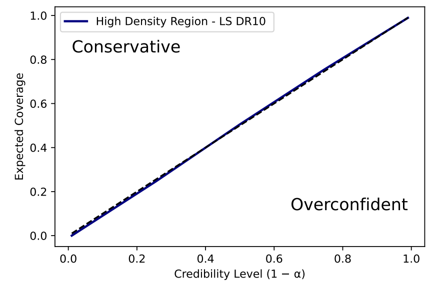

Figure 7. Coverage test results, where the blue solid line represents the resulting curve for our estimated PDFs and the black dashed line represents the ideal case of perfectly calibrated PDFs in terms of credibility.

brated, or underconfident (Hermans et al. 2022). The coverage diagnostic was executed by selecting the pair (spec-z, PDF) and sampling 1000 values from the photo-z's PDF, thereby generating a frequency distribution with 1000 bins spanning the range from 0 to 1.

First, a credible region is defined for the estimated distribution using the Highest Density Regions (HDR). This region represents the smallest area that contains at least $100(1-\alpha)\%$ of the mass of the inferred photo-z distribution, establishing an interval for a given credibility level $(1-\alpha\ )$. The expected coverage is the frequency with which the true parameter (spec-z) value falls within this highest density region, essentially indicating how often it falls inside the calculated interval. If our model produces well calibrated distributions we expect the spec-z value should be contained inside the calculated interval from the $(1-\alpha)$ HPD region of the estimated distribution exactly $(1-\alpha)\%$ of the time. If the coverage probability is less than the $(1-\alpha)$ credibility level is the sign of under-estimation of the PDF's variance, and it could lead to unreliable approximations since it excludes physical values of photo–z . Conversely, if the coverage probability is larger than the $(1-\alpha)$ credibility level, then this indicates that the estimated PDFs are over-estimating their variance, in average.

The middle panel of Fig. 6 show the odds distribution for the same set of galaxies present on the point-estimate analysis, while the left panel illustrates the PIT distribution for a subset of these objects satisfying the constraint $odds > 0.7$. The latter exhibits a marginally concavity at lower values of PIT and a negative slope throughout higher values. Although not pronounced, this features collectively suggest a tendency towards positive bias, implying a slight overestimation of the most probable photo–z 's values. As discussed above, there is trending in the median bias curve (Fig. 4) of overestimating photo–z . This overestimation effect is further illustrated in Figure 5, particularly evident around photo–z equal to 0.48. Despite this anomaly at the higher redshifts, the scattering of photo–z remains well behaved, as expected. Consequently, the inclination of the PIT distribution might be attributed to a localized systematic bias rather than indicating any inherent global bias within the model. Although this could impact the analysis, it is worth stressing that our Bayesian analysis performs a bias correction based on this curve in the same manner as previous work by Alfradique et al. (2024). A similar shape of PIT distribution was also encountered by D'Isanto, A. & Polsterer, K. L. (2018) using MDN to estimate redshifts from images in the Sloan Digital Sky Survey Data Release 9.

The odds distribution shows a gradual increase in frequency with the odds values. This result suggest the absence of a systematically overconfident or overdispersed PDFs, indicating a high level of confidence in the estimated PDFs. From Fig.7, a perfectly calibrated estimated distribution aligns with a diagonal line (dashed black line), indicating an expected coverage probability matching the credibility level. We notice that the expected coverage curve matches the diagonal, indicating that our model produces well calibrated photo–z PDFs. This linear relation confirms the reliability of our methodology.

In summary, our methodology estimates photo–z with an accuracy comparable to those publicly available in LS-DR10, while also generating well-calibrated PDFs. It is noteworthy that our approach can predict redshifts in objects with detections in all g,r,i,z bands, and in any combination of three bands containing detections in g band.

B. METHOD

Here, we outline the statistical methodology known as the dark siren approach, initially introduced in Chen et al. (2018) and subsequently modified by (Soares-Santos et al. 2019; Palmese et al. 2020; Palmese et al. 2023; Alfradique et al. 2024). The method consists of using the Bayesian formalism to infer the Hubble constant (H_0) parameter through the gravitational wave detection data $d_{\rm GW}$ and the $d_{\rm EM}$ set of photo-z measurements of the possible host galaxies made using the deep learning algorithm (see Appendix A). As the GW and EM measurements are done independently, the joint GW and EM likelihood can be defined as the product of the two individual likelihood, $p\left(d_{\rm GW},d_{\rm EM}|H_0\right) = p\left(d_{\rm GW}|H_0\right) p\left(d_{\rm EM}|H_0\right)$.

From the Bayesian framework, the H_0 posterior of one GW event can be written in the final form as

$$p(H_0|d_{\rm GW},d_{\rm EM}) \propto \frac{p(H_0)}{\beta(H_0)}\sum_i \frac{1}{\mathcal{Z}_i}\int {\rm d}z\,{\rm d}\Delta z\, p(d_{\rm GW}|d_L(z,H_0),\hat{\Omega}_i)p_i(d_{\rm EM}|z,\Delta z)\; p(\Delta z)\frac{r^2(z)}{H(z)}\,. \tag{B5}$$

where $p(H_0)$ is prior on H_0 which we assume to be flat over the range $[20,140]$ km s^{-1} Mpc^{-1}, $\beta(H_0)$ is the normalization factor that describes the selection effects in the measurement process, $r(z)$ is the comoving distance, $H(z) = H_0\left(\Omega_m(1+z)^3+1-\Omega_m\right)^{1/2}$ is the Hubble parameter in a Flat ΛCDM model, $\mathcal{Z}_i = \int p(d_{\rm EM}|z_i)r^2(z_i)/H(z_i)\,{\rm d}z_i$ are evidence terms that normalize the posterior, and $p(\Delta z)$ is the prior on the photo-z bias that we measure from our photo-z validation sample (see Fig. 4 in appendix A). In the following paragraphs we will detail how the function $\beta(H_0)$ was computed. The GW and EM likelihoods are written as, respectively:

$$\begin{gathered} p(d_{\rm GW}|d_L(z,H_0),\hat{\Omega}_i) \propto \frac{p(\hat{\Omega}_i)N(\hat{\Omega}_i)}{\sqrt{2\pi}\sigma(\hat{\Omega}_i)}\exp\left[-\frac{(d_L-\mu(\hat{\Omega}_i))^2}{2\sigma^2(\hat{\Omega}_i)}\right],\\ p(d_{\rm EM}|z,\Delta z) = \prod_i p(\bar{z}_i|z,\Delta z), \end{gathered} \tag{B6}$$

where we explicitly consider the dependence of cosmological parameters on the luminosity distance, as measured by GW, and the solid angle $\hat{\Omega}_i$ corresponding to each observed galaxy i. Following Singer et al. (2016), the GW likelihood is approximated by a Gaussian function. The EM likelihood is the product of the probability distribution function of the photometric redshift $\bar{z}_k$ for each k galaxy, where we consider the correction of the shifted in redshift for the photo-z biases, Δz, in the data. This phenomenon arises due to the absence of a uniform distribution in redshift or color beyond a certain magnitude or color selection threshold, which makes the deep learning algorithm begin to oversample the redshift around the distribution peaks, thereby introducing systematic biases. The individual H_0 posterior distributions found in this work are presented (in colors) in Fig. 8.

The selection effects were computed based on the criteria outlined in Chen et al. (2018) and Gray et al. (2020). This involves the joint gravitational wave-electromagnetic likelihood, which is marginalized over all conceivable GW and EM data. Assuming that the events are isotropically distributed on a large scale, this term can be written in its compact form as:

$$\beta = \int_0^{z_{\rm max}} p_{\rm sel}^{\rm GW}\left(d_L(z,H_0)\right)p(z)\,dz \tag{B7}$$

where $p_{\rm sel}^{\rm GW}$ is the probability of a GW event at a given luminosity distance d_L be detected, $p(z)$ is the galaxy catalog distribution, and $z_{\rm max}$ is the maximum true redshift at which we can detect the host galaxies. The function $p_{\rm sel}^{\rm GW}$ is equal to one if all events in a given redshift z satisfy the detection condition, i.e. the detector network SNR is > 12 and the localization volume satisfies our selection criteria ($A_{50\%} < 1000\,{\rm deg}^2$ and $d_L < 1500\,{\rm Mpc}$), and zero if none of the events located in a redshift z satisfies those conditions. Therefore, $p_{\rm sel}^{\rm GW}$ function is, in sum, an efficiency curve, i.e. a smooth function that falls from 1 to 0 over the redshift range for each H_0 value.

The selection function was modeled by simulating 30,000 BBH mergers using BAYESTAR software (Singer & Price 2016; Singer et al. 2016; Singer et al. 2016) with the frequency domain approximant IMRPhenomD. We simulate the BBH mergers assuming that they follow a power law plus peak mass distribution with the same parameters as described in Bom & Palmese (2023), the spins distributions follow a uniform distribution between $(-1,1)$, and a uniform distribution in comoving volume assuming a dependency with the merger rate evolution described by the Madau-Dickinson star formation rate (Madau 2014) and fixing the Planck 2018 cosmology for twenty different H_0 values within our prior range. We assumed that the Legacy galaxies catalog is complete up to the known absolute magnitudes for each of the GW events in the analysis, which correspond to a maximum redshift of $z_{\rm max} = 0.65$. All the 30,000 injected events passed by the matched-filter analysis, where we computed the signal-noise-ratio (SNR) assuming the O4 sensibility curve as published by LIGO in Document P1200087 (https://dcc.ligo.org/LIGO-T2000012/public) for the Hanford and Livingston LIGO network detectors. The network SNR defines the detection condition above 12 and at least 2 detectors have a single–detector SNR above 4. A Gaussian noise was added in all the measurements. Lastly, the BAYESTAR skymaps were reconstructed for each detection event, where we assume a luminosity distance that follows $\propto d_L^2$.

The last selection cut that we should consider is that events serving as dark sirens have at least 70% of their 90% CI comoving volume covered by the Legacy survey. As noted by Palmese et al. (2023), this selection cut can be ignored since the GW antenna

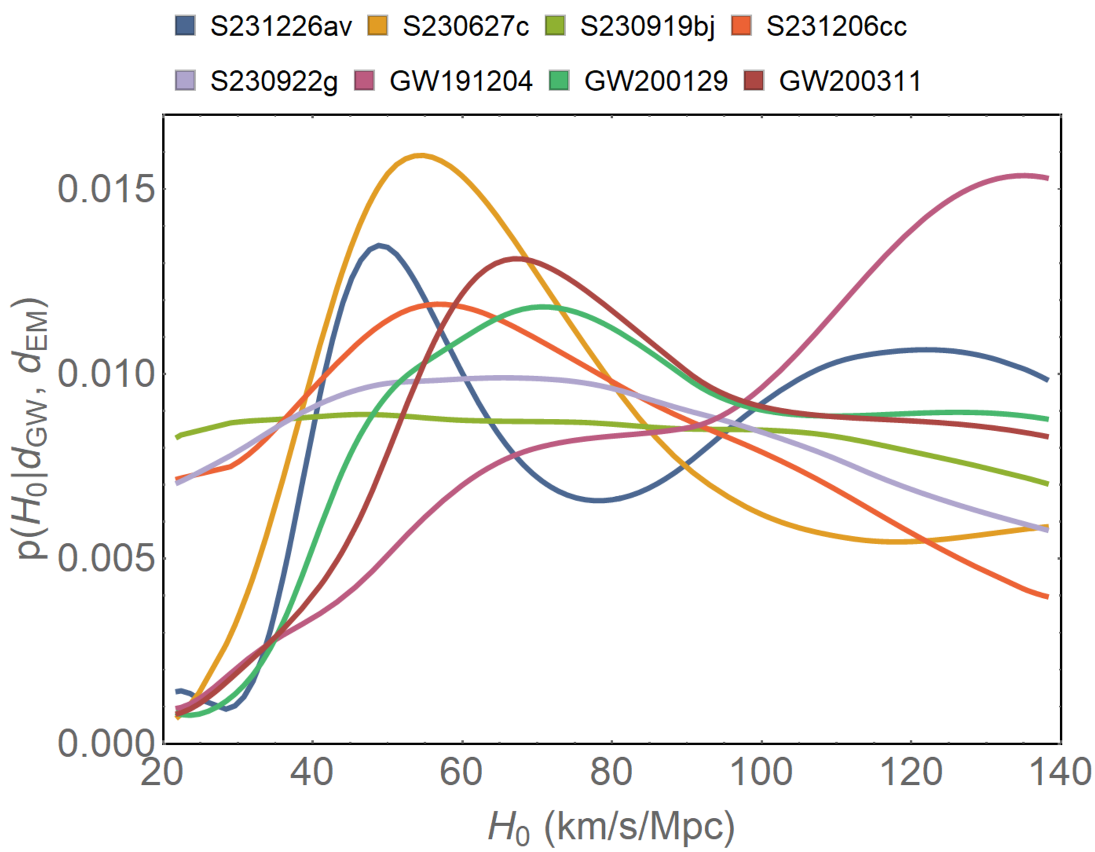

Figure 8. The Hubble constant posterior distributions for each dark siren considered in this work, which were found using the galaxies from Legacy survey.

pattern is not correlated to the survey sky footprint and we do not expect an H_0 dependence. Here, we also ignore this selection effect.

C. COMPARISON WITH H_0 CONSTRAINTS FROM POINT ESTIMATES OF PHOTOMETRIC REDSHIFTS

During this work, the H_0 posterior calculation was performed using the full photo-z PDFs estimated through the deep learning method, described in Appendix A. In this appendix we will analyze the effect on the H_0 posterior originated by the choice of the galaxies photo-z PDFs, the photo-z PDF estimated by the MDN technique described in Appendix A, which we will call the full photo-z PDF, or a Gaussian approximation, where we will approximate each full photo-z PDFs to a Gaussian function whose mean coincide with the peak value and the value of the standard deviation is equal to that found with the full photo-z PDF. In the sense that the Gaussian approximations were constructed with the same peak and standard deviation as that found with the PDFs estimated with the MDN. We include the results of the public Legacy [5] in the figs. The top panel of Fig. 9 shows the posterior of H_0 found considering the full photo-z PDF (blue solid curve) and the Gaussian approximation (black dashed curve for the full PDF, this work, and the Legacy DR10.1 public results in gray dashed curve), and the residuals curves are shown on the bottom panels. Comparing the results of the posterior of H_0 generated considering the Gaussian approach (black dashed curve) and the full photo-z PDF, we note that the results are in agreement at the percentage level for the more distant events, where the effect is suppressed by marginalization over a larger number of galaxies due to its larger localization volume. The effect of the use of the Gaussian approximation is more evident for the events S230627c, GW200129 and GW191204, which present a discrepancy of $\approx 4\%, 16\%$, and 40%, respectively, in the peak distribution and can reach values $> 13\%$ at the ends. As expected, the Gaussian approximation makes the peak of the H_0 posterior wider, implying a less precise H_0 measurement. This result is a consequence of the smoothing of overdensity regions in the photo-z distribution. Although the results do not indicate that the choice of photo-z PDF leads to disagreement H_0 measurements, future dark siren measurements can reach a level of precision where the differences

[5] The photo-z were computed using the Random Forest technique with data from the Legacy DR10.1, these results are available in https://www.legacysurvey.org/dr10.

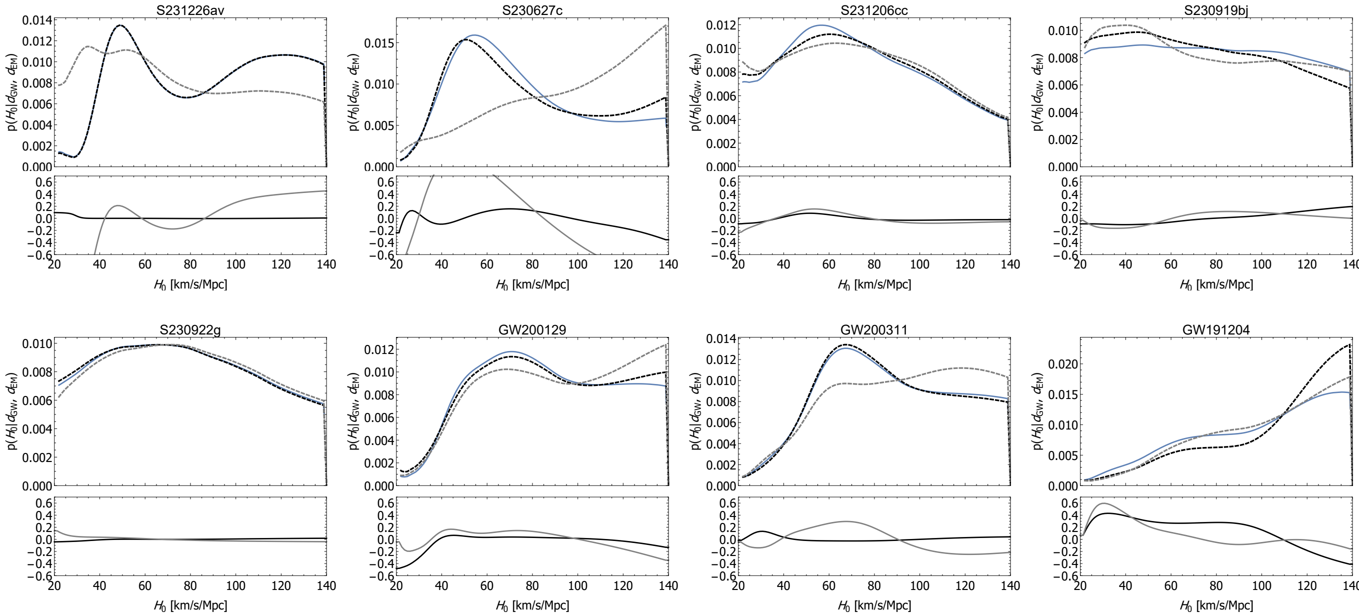

Figure 9. Comparison between the Hubble constant posterior distributions found by using the full galaxies redshift PDFs (blue solid line) and a Gaussian approximation (dashed line: (1) black color using the means and standard deviation of the full PDFs, and (2) gray color using the photo-z measurements made by the public Legacy DR10.1) for all the dark sirens study here. The bottom panels show the residuals value between the two curves shown on the top panel, calculated as $2 \times (\text{solid line} - \text{dashed line})/(\text{solid line} + \text{dashed line})$.

between these results may be statistically significant. We can also see that the Legacy public results leads to, in most events, a less restrictive H_0 posterior compared to those achieved through the full photo-z PDFs, leading to a increase of up to 6% in the H_0 uncertainty.

As shown by Soares-Santos et al. (2019), the choice for a wider redshift cut causes galaxies to be added in deeper redshift, whose photo-z PDF has the a significant Gaussian tails at high redshift, which implies a increase of the H_0 posterior at high H_0. However, the full photo-z PDF is able to reduce (the residual values, between the posterior considered a LIGO/Virgo luminosity distance posterior of 90% and 99.7%, at the high-H_0 end reduced by $\approx 25\%$ compared to the result achieved with the Gaussian approximation) this dependence of the H_0 posterior behavior with redshift cut, which evidences its advantage against the Gaussian approximation in performing a measure of H_0 free of any systematic imposed by assumptions made in the methodology.

In Fig.10 we present the photo-z PDFs estimated by the deep Learning method (dark blue curve) and the Gaussian approximation (dashed curve in light blue) for galaxies at different redshifts. We notice that the full photo-z PDF has a tendency to be narrower than the Gaussian approximation, which favors a more precise measurement of H_0. In addition, we see that the difference between the estimated PDF and the Gaussian approximation becomes more evident in higher redshifts. This highlights the importance of the MDN technique mainly for galaxies located in deep redshifts, that are already used for the LIGO/Virgo GW dark sirens and will be increasingly indispensable for the coming years where it is expected to observe increasingly distant GW events. We can also observe that the full photo-z PDFs become increasingly non-Gaussian in deeper redshifts. In order to quantify this non-Gaussianity in Fig. 11 we present the relation between the kurtosis (fourth standardized moment $\tilde{\mu}_4$, that is presented in the left panel) and the skewness (third standardized moment $\tilde{\mu}_3$ shown in the right panel) of the full PDFs with $z_{\rm phot}$, where we compute the mean (solid line) and the standard deviation (shaded region) of this quantities in photo-z bins of size 0.05. The values of kurtosis and skewness are distinct from the values of a Gaussian distribution (kurtosis and skewness equals to 3 and 0, respectively) in almost the entire $z_{\rm phot}$ interval. We note a tendency of kurtosis values to be lower with increasing $z_{\rm phot}$, which indicates the enlargement of PDFs implying in less precise $z_{\rm phot}$ measures. This behavior describes the already expected difficulty of precise measurements of distant galaxies. The skewness results indicate that the full PDFs are right-skewed in low $z_{\rm phot}$, being reduced with the increase of $z_{\rm phot}$ until reaching the value of zero (indicating a symmetric PDF) in $z_{\rm phot} \approx 0.5$, in which from this value begins to be left-skewed.

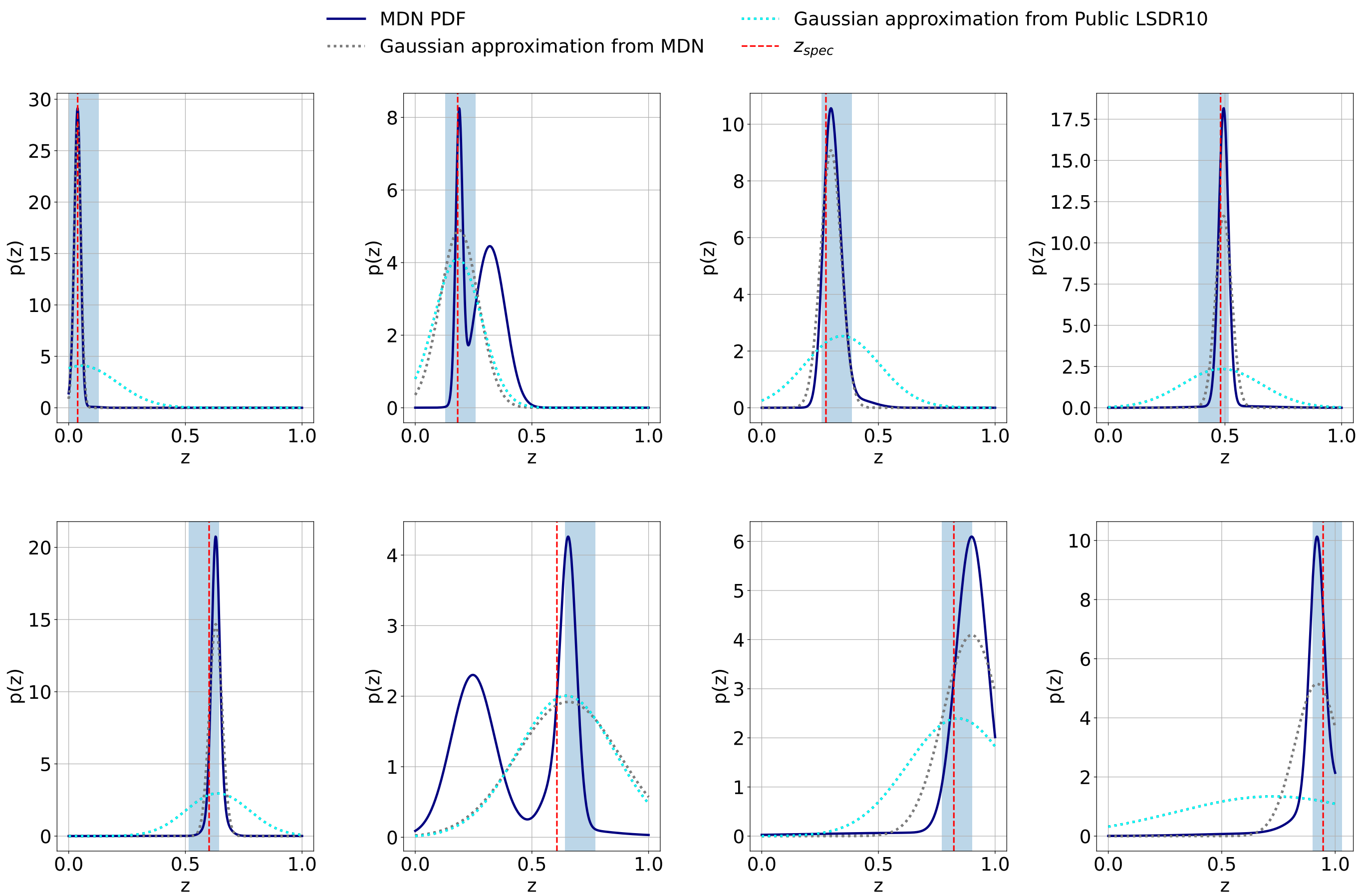

Figure 10. Photometric redshift PDFs estimated with the MDN technique (dark blue line), Gaussian approximation of MDN photo-z PDF (dashed curve in cyan), and Gaussian approximation using the photo-z estimated by the Public LSDR10 (dashed curve in gray) for different redshifts (increases from left to right). The vertical dashed red line represents the corresponding $z_{\rm spec}$ value.

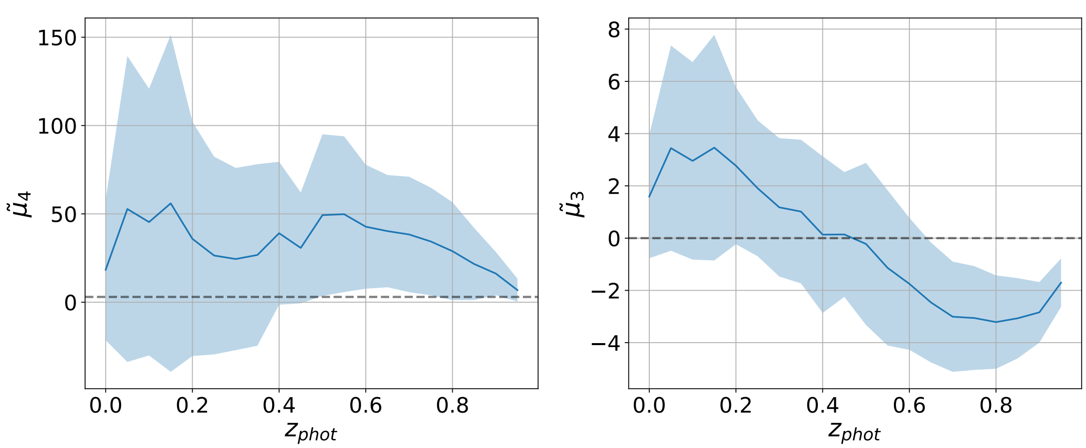

Figure 11. Kurtosis (left panel) and skewness (right panel) values of the photo-z PDFs, estimated with the MDN technique, as a function of $z_{\rm phot}$. The solid line represents the mean value of each quantity and the shaded region represents the respective standard deviations computed in photo-z bins of size 0.05.