Understanding Combined Results From Multiple GW Searches Using Information Theory

Oleksandra Lukina
University of South Dakota

Mentor: Derek Davis
LIGO, California Institute of Technology
(Dated: October 9, 2023)

Final Report
LIGO SURF 2023

Determining whether a gravitational wave (GW) signal is of astrophysical origin or is caused by terrestrial noise still presents a challenge to the GW community. Current searches estimate the significance of events by calculating the false alarm rate (FAR) and p_{astro}, but these results are limited to a single search pipeline. In this work, we investigate three different methods of combining information from multiple GW searches and calculating a joint p_{astro}, including the application of the Bonferroni correction to individual FARs, finding a harmonic mean FAR, and using maximum individual p_{astro} as a measure of event significance. Using these approaches, we compare the effectiveness of different combinations of searches using the language of information theory and show that purity of current Gravitational-wave Transient Catalogs (GWTCs) is likely overestimated.

I. INTRODUCTION

Since the first detection of gravitational waves (GWs) [1], the total number of GW candidates reported by LIGO-Virgo-KAGRA (LVK) Collaboration reached 90 [2] and continues to grow [1]. At the same time, determining whether a certain signal has astrophysical origin or is caused by terrestrial noise still remains a challenge, which leads to the uncertainty in the number of detected compact mergers [3].

Noisy local environments that are difficult to model, observations with multiple detectors, and inability to shield the instruments from GW signals result in the dependence of the estimated event significance on a search analysis [4]. Two main approaches that are used to search for events include matching signals to the compact binary coalescence waveform templates and searching for transient signals across the network of detectors with minimal modeling. Variations of the first method are used in PyCBC [5], GstLAL [6], MBTA [7], IAS [8], and OGC [9] pipelines, and the second method is used in the cWB pipeline [2]. There are several technical differences between search pipelines, so they result in different estimations of event significance, which often leads to contradictory conclusions for the same detector data. However, all pipelines are designed to search for the same signals, so their results are not fully independent and should correlate [10].

In order to estimate significance of an GW candidate, Gravitational-wave Transient Catalogs, e.g., GWTC-3 [2], report three quantities, namely signal-to-noise ratio (SNR), false alarm rate (FAR), and p_{astro}, for each search pipeline that detected the event. False alarm rate, usually measured in yr^{-1}, is a rate of coincidence triggers that occur due to noise alone and have SNR equal or higher than a certain value. As a result, to confirm the presence of a signal, one must show that the probability to obtain the observed event in a dataset that only contains noise is smaller than a given threshold [4].

On the other hand, p_{astro} is defined as a probability that a GW candidate has astrophysical origin and is not caused by terrestrial noise. It is calculated by combining the rates at which triggers – outputs of a search pipeline – are generated by both astrophysical and noise sources, i.e., both false and true alarm rates [11].

From the statistical point of view, the problem of analysing results from multiple search pipelines is a multiple-comparison procedure (MPS). In this framework, controlling errors in p_{astro} values and false alarm rates in presence of multiple searches is controlling the family-wise error rates and false discovery rates. Statistical and medical research shows that unguarded use of single-inference results and failure to apply appropriate corrections when pursuing multiple inferences greatly increases false positive rate and jeopardizes sensitivity to detect true signals [12, 13]. Currently, GWTC catalogs, use the maximum of single-pipeline p_{astro} and FAR values as a measure of event significance, which is an example of such an unguarded approach. As a result, we are interested in evaluating the effectiveness of this method in comparison with other statistical solutions to the MPS problem in the context of GW search pipelines.

For the purpose of this work, we analyze the search results only in terms of the FAR and p_{astro} distributions they produce. Consequently, we do not take into account the differences in FAR calculation and methods for esti-

[1] https://observing.docs.ligo.org/plan

mating noise properties across the network of pipelines. This data analysis approach can be considered to be an application of information theory with the number of events passing FAR and p_{astro} thresholds as a macroscopic parameter.

II. METHODS

The goal of this project is to investigate different ways of combining information from multiple GW search pipelines and analyze what new knowledge these combinations provide about observed compact mergers. In addition, we are interested in analysing the contributions of the IAS and OGC pipelines, developed outside of the LVK collaboration, to the results of the internal pipelines.

In order to do this, we introduce three different methods of calculating a combined measure of significance of GW events for any number of pipelines using only their FAR distributions:

- **Trials**: Find the combined FAR by applying the Bonferroni correction (trials factor) [14] to FARs from individual pipelines corresponding to the same trigger.

$$FAR_{1...N} = \frac{1}{N} \min\left(FAR_1, FAR_2,\ \dots\ , FAR_N\right) \quad (1)$$

- **Harmonic**: Find the combined FAR by calculating a harmonic mean of individual FARs [13].

$$\frac{1}{FAR_{1...N}} = \frac{1}{N}\left(\frac{1}{FAR_1} + \frac{1}{FAR_2} + ... + \frac{1}{FAR_N}\right) \quad (2)$$

- **Max**: Calculate p_{astro} values for individual pipelines from their FAR distributions. Assign the highest p_{astro} calculated for a certain trigger to be the combined p_{astro}. This method is used in GWTC-3 catalog [2].

$$p^{1...N}_{astro} = \max\left(p^1_{astro}, p^2_{astro},\ \dots\ , p^N_{astro}\right) \quad (3)$$

The first two methods result in new FAR distributions that incorporate information from all pipelines and can be the basis for calculating a combined p_{astro}. On the other hand, the third method uses original FAR distributions to calculate separate p_{astro} values, which then need to be combined. Since search pipelines are designed to look for the same modeled waveforms, we expected their results to be correlated. According to statistical studies, in case of dependent tests the Harmonic method is expected to have a statistical advantage and reduce the rate of false positive detections [13].

After calculating combined FARs from the Trials and Harmonic methods, we analyse how many events from the first part of the third observing run (O3a) and the LVK injection set of simulated events pass combined FAR thresholds. In order to do this we use the following data:

- **LVK data**: The set of FARs of all triggers that were detected by at least one pipeline among PyCBC (highmass and all-sky treated as two separate pipelines), GstLAL, or MBTA during the O3a obsering run [15], accesed via [2].

- **IAS data**: The set of FARs of events detected by the IAS [16] pipeline, data accessed via [3].

- **OGC data**: The set of FARs of events detected by the OGC [9] pipeline, data accessed via [4].

- **Injection data**: The set of FARs for the O3a injection set that includes five LVK searches, namely PyCBC (BBH and hyperbank treated as two separate pipelines), GstLAL, MBTA, and CWB that contains 512431 injections. Accessed via [5].

Next, we use combined FAR distributions for Trials and Harmonic methods and original FAR distributions for Max method to calculate p_{astro} values using the FGMC method [17]:

$$p(\Lambda_s, \Lambda_n|x) = \prod[\Lambda_s f(x_i) + \Lambda_n b(x_i)]e^{-(\Lambda_s+\Lambda_n)}, \quad (4)$$

where x – ranking statistic (e.g., FAR), Λ_s – number of signal counts; Λ_n – number of noise counts; $f(x_i) = f(FAR_i)$ – foreground (signal) model, determined from the injection set; $b(x_i) = b(FAR_i) = 1/FAR_i$ – background model, estimated analytically.

We use this equation to fit the empirical FAR distribution for confidence intervals of Λ_s and Λ_n using `log_likelihood` method from `bilby` Bayesian inference Python library. Then, we calculated the range of p_{astro} values using the minimum Λ_s and maximum Λ_n for the lower end of the p_{astro} confidence interval and the maximum Λ_s and minimum Λ_n for the upper end as follows:

$$p_{astro}(FAR) = \frac{\Lambda_s f(FAR)}{\Lambda_s f(FAR) + \Lambda_n b(FAR)} \quad (5)$$

By doing this calculation, we obtain combined p_{astro} values for Trials and Harmonic methods and apply equation (3) to find the combined value for the Max method. This allows us to compare the effectiveness of all three

---

[2] https://zenodo.org/record/5759108
[3] https://github.com/seth-olsen/new_BBH_mergers_O3a_IAS_pipeline
[4] https://github.com/gwastro/4-ogc
[5] https://zenodo.org/record/5117799

approaches by comparing their combined p_{astro} results. First, we compare the number

$$N = |\{\, p_{astro} \mid p_{astro} > 0.5 \,\}| \tag{6}$$

of events whose largest p_{astro} values consistent with the uncertainty pass the p_{astro} threshold for the real O3a data and the set of injections using the three methods. Then, using real data, we analyze the purity of results for each method using the following equation:

$$purity = \frac{\sum_{i=1}^{N} p^{i}_{astro}}{N}, \tag{7}$$

which is a sum of p_{astro} values for events that passed a threshold of 0.5 divided by the number of those events. The sum of p_{astro} values can be interpreted as the number of real events present in the set, which makes the calculated fraction a measure of purity of the catalog. However, since purity estimates depend on the method of calculating p_{astro}, if the p_{astro} calculation is performed incorrectly, purity results will also be incorrect.

III. RESULTS

A. False Alarm Rates

By applying methods 1 and 2 to the injection set, we found that the harmonic mean combination of all pipelines recovers about 1.1% more events with $FAR < 1\ yr^{-1}$ than the combination that uses the trials factor. Since the injection set consists of simulated events and does not allow for false positive detections, the higher number of recovered signals shows that the harmonic mean method has a slightly better performance. However, this difference was not noticeable for the real signals ($N < 50$), and the two methods resulted in effectively the same values for O3a events. As a result, we will provide only the harmonic mean combined values for most of the analysis in this section to avoid redundancy.

Using the equation (2), we calculated combined FARs for all events that were detected by at least one of the pipelines included in each of the following datasets:

1. O3a data for GstLAL (g), PyCBC (p), PyCBC BBH (b), MBTA (m), IAS (i), and OGC (o).

2. O3a data for GstLAL (g), PyCBC (p), PyCBC BBH (b), and MBTA (m).

3. Injection set data for GstLAL (g), PyCBC (p), PyCBC BBH (b), MBTA (m), and cWB (c)[6].

[6] PyCBC = hyperbank = all-sky, PyCBC BBH = highmass.

Then, we calculated the number of events that pass the threshold of combined $FAR < 1\ yr^{-1}$ for each combination of different number of pipelines. Depending on a combination, joint searches detected between 24 and 40 events in dataset 1, 24-35 events in dataset 2, and 71763-96173 events in dataset 3 as compared to 22-35 events detected by each individual pipeline in datasets 1 and 2 and 18260-77289 events in the injection dataset 3.

Based on these calculations, we identified the combinations of pipelines that detect the most events for each number of pipelines in a group and presented these results in a form of decision trees shown in Fig. 1. For example, among the combinations of two pipelines used in dataset 1, the highest number of events (40) was detected by the combination of GstLAL (g) and IAS (i) searches, which is labeled as "gi". Interestingly, the maximum number of detected events decreases as we add more searches for real data (datasets 1 and 2) and increases for the injection set (dataset 3). This difference can be explained by the presence of false positive detections in the real data but not in the simulated data, which will be discussed in more details in Conclusions.

In the top part of Fig. 1, we notice that both the GstLAL (g) pipeline and the combination of all searches (gibmpo) detect the same number of signals, which is 35. 31 of them are the same, but there are four signals that are detected only by GstLAL and not by the combination (GW190426_152155, GW190431_023648, GW190731_140936, GW190917_114630), and different four signals that are detected by the combination and not by GstLAL alone (GW190413_134308, GW190514_065416, GW190725_174728, GW190925_232845).

The next question that we wanted to address is whether the trend of recovering more injected events with addition of more pipelines, observed in the bottom part of Fig. 1, is true for any threshold value. In Fig. 2, we compare the number of recovered injections for the same five combinations of searches (g, gp, gpb, gpbm, and gpbmc) and for thresholds from 10^{5} to $10^{-4}\ yr^{-1}$. The graph shows that adding more pipelines leads to more detected events for thresholds larger than $10^{-4} yr^{-1}$, but for very low thresholds the difference becomes negligible. Moreover, the curves for "gpbm" and "gpbmc" almost perfectly overlap, which illustrates that the addition of the fifth (cWB) pipeline does not lead to any significant increase.

B. p_{astro}

In this section, we compare combined p_{astro} results for injection sets calculated using the three methods described earlier. The results of this comparison are shown in Fig. 3, where we can see that all three combined searches find more events with $p_{astro} > 0.5$ than any individual pipeline included in the combination. The Max method detects about 2.3% more events than the Tri-

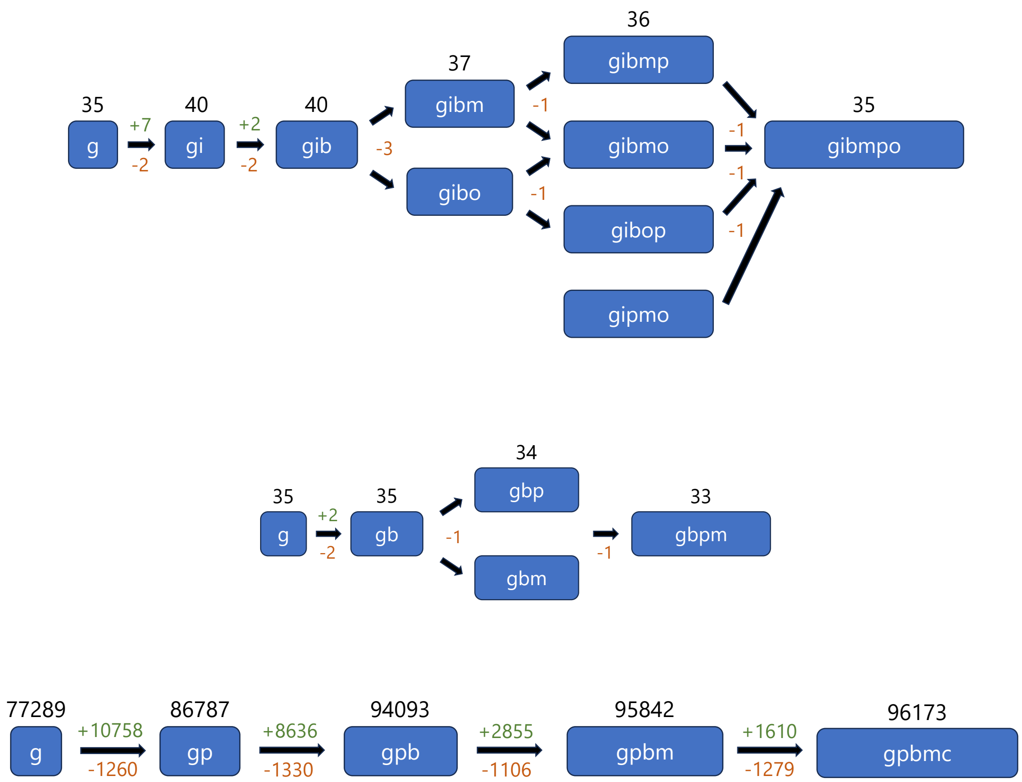

FIG. 1: Combinations of pipelines that detect the most events for each number of searches. GstLAL (g), PyCBC (p), PyCBC_BBH (b), MBTA (m), cWB (c), IAS (i), and OGC (o). Top: Combinations of LVK, IAS, and OGC pipelines based on O3a data. Middle: Combinations of LVK pipelines based on O3a data. Bottom: LVK pipelines based on the injection set. Numbers above the boxes show the number of candidates passing the threshold of $FAR < 1\ yr^{-1}$, and numbers above and below the arrows indicate the number of events lost and gained due to the addition of an extra pipeline to the calculation of the combined FARs.

als method, and the Trials method detects 0.2% more results than the Harmonic method. However, the differences between these combinations is much smaller than the advantage each of them gives in comparison to any of the individual pipelines.

For real events, however, the same comparison leads to very different results. In Fig. 4, the bars correspond to the number of events detected by each search or a combination of searches, and we see that the Trials and Harmonic methods detect less events than the GstLAL pipeline alone. Moreover, the Max methods detects 58-62% more events than the first two combination methods.

At the same time, in case of real events, we can calculate the purity of each search using equation (7), and the results are shown above each bar. Blue (darker) portions show the sum of p_{astro} values or the number of real events we expect to see among all detected events, and the gray (lighter) parts show the remaining number $N - \sum p_{astro}$, which corresponds to noise. From this result, we can see that, although the Max method detects significantly more events, it also leads to lower purity of the results. At the same time, all individual pipelines as well as Trials and Harmonic combinations have purity estimates close to 90%.

Finally, to characterize the Max method, we calculate the purity of the 76 events it detected using the sum of p_{astro} values found with the Harmonic method. The result is 69%, which is noticeably lower than the 85% purity from Max method, as shown in Fig 5.

IV. CONCLUSIONS

In this work, we described and compared three different methods of combining FAR and p_{astro} results from multiple pipelines. We found that GstLAL detects the most events among individual pipelines for injections and O3a data both in terms of FAR and p_{astro}. At

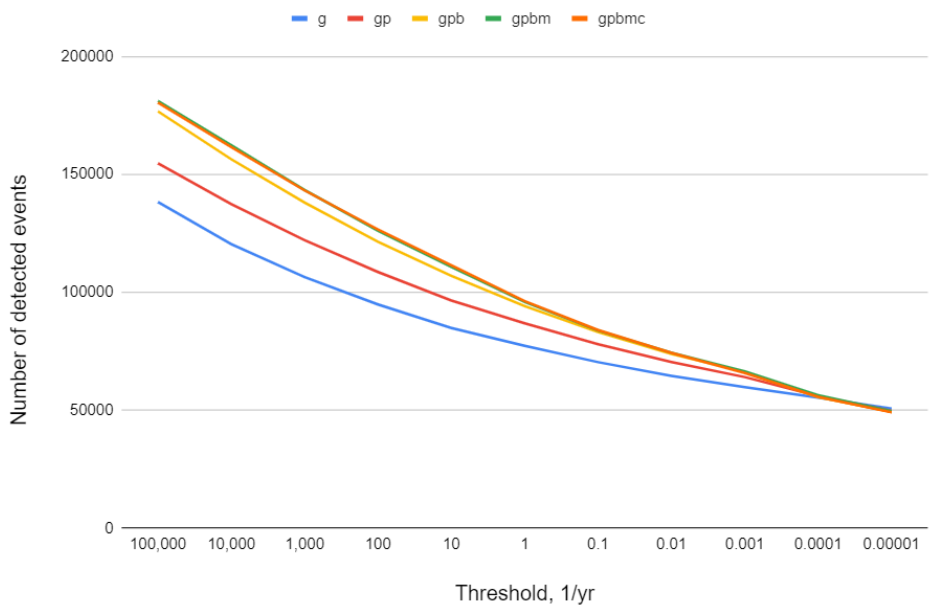

FIG. 2: Number of injected events recovered with combined FAR below a certain threshold depending on a group of combined searches and a threshold value in yr^{-1}. g – GstLAL, gp – GstLAL and PyCBC, gpb – pg with PyCBC_BBH, gpbm – pgb with MBTA, gpbmc – gpbm with cWB. Adding more pipelines leads to more detected events for thresholds larger than $10^{-4}yr^{-1}$.

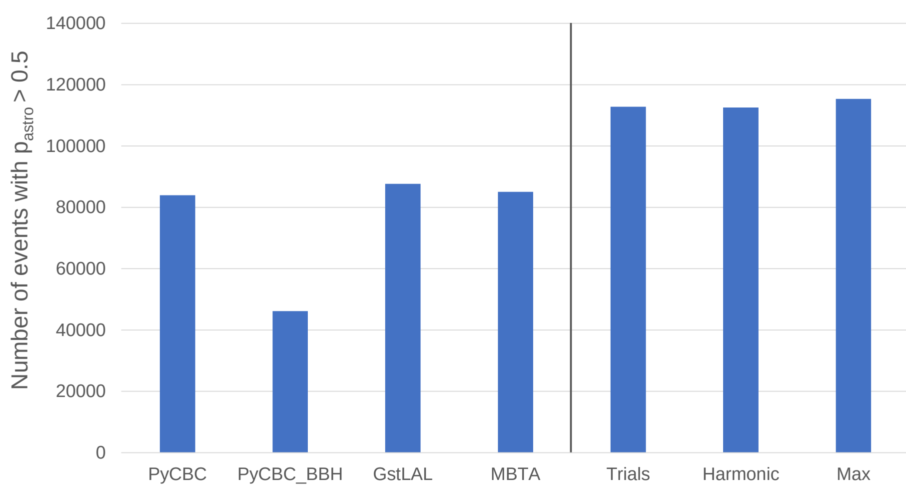

FIG. 3: Number of injections recovered with $p_{astro} > 0.5$ by four LVK pipelines and three ways to combine their results, namely Trials, Harmonic, and Max methods. The figure illustrates that combined searches detect more injected events than individual ones.

the same time, results from the injection set show that combinations of pipelines detect significantly more events that pass FAR and p_{astro} thresholds than any individual pipeline.

On the other hand, combining results from more pipelines for O3a data does not lead to any increase in the number of events passing the FAR threshold, and only one of the combination methods found more events passing the p_{astro} threshold than GstLAL. This can either mean that combining pipeline results using our methods

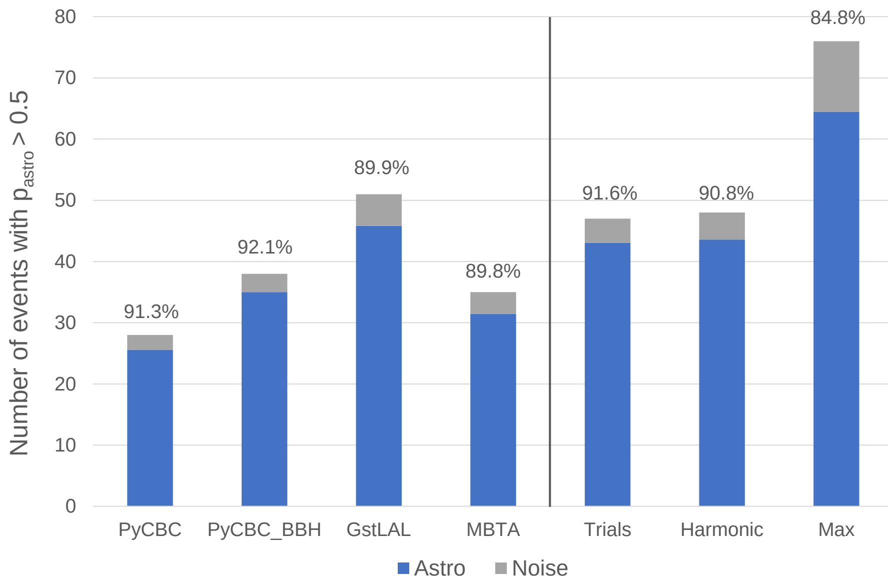

FIG. 4: Number of real events detected with $p_{astro} > 0.5$ for four LVK pipelines and three ways of combining them. The purity of each result is shown above each bar, and colors indicate the number of astrophysical and noise events expected in each set of detections. This figure shows that Max method detects the most events, but also leads to lower purity of the results.

is not effective for available data or that the smaller number of signals is more accurate and possibly eliminates the false positive detections present in single-pipeline results.

Among three different ways of combining FAR distributions, applying the trials factor and finding a harmonic mean give similar p_{astro} results. The method of finding the maximum of individual p_{astro} values finds 2.3-2.5% more events for simulated data and 58-62% more events for real data than the other two methods. However, while the estimated purity for Trials and Harmonic methods stays close to 91-92% and is similar to the estimates for individual pipelines, the purity of the Max search might be as low as 69%. Since Max method is used in current LVK p_{astro} analysis, this result suggests that purity of the GWTC catalogs is likely overestimated.

The results obtained with injected events are not subject to false positive detections, so all three combinations of pipelines lead to an increase in true positive rates with the Max method having the highest and Harmonic method having the lowest values. At the same time, purity results for real events can be used to qualitatively analyse the false positive rates of the same combinations. The Trials method showed the highest purity (91.6%), followed by Harmonic method (90.8%), and the Max method showing the lowest result (84.8%), while single-pipeline purity ranging from 89.8 to 92.1%. This result is consistent with the statistical fact that using single-inference results in the multiple-comparison problem increases the false positive rate [12].

In gravitational-wave searches, an increase in true positive rates leads to the detection of more events, but an increase in false positive rates can lead to false-alarm alerts for astronomical follow-up as well as an exaggerated picture of black hole and neutron star populations. The results of this work show the need of combining results from multiple pipelines while also controlling the family-wise error rates, false discovery rates, and purity. Our analysis indicates that the Bonferroni correction applied to FARs might be the simplest working solution to this problem, but further investigations are needed to find the most statistically justified approach, especially once more GW detections are available for statistical studies.

V. ACKNOWLEGEMENTS

The authors gratefully acknowledge contributions and support of the California Institute of Technology LIGO Laboratory, California Institute of Technology Student-Faculty Programs, and the National Science Foundation.

This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the National Science Foundation, and operates under co-

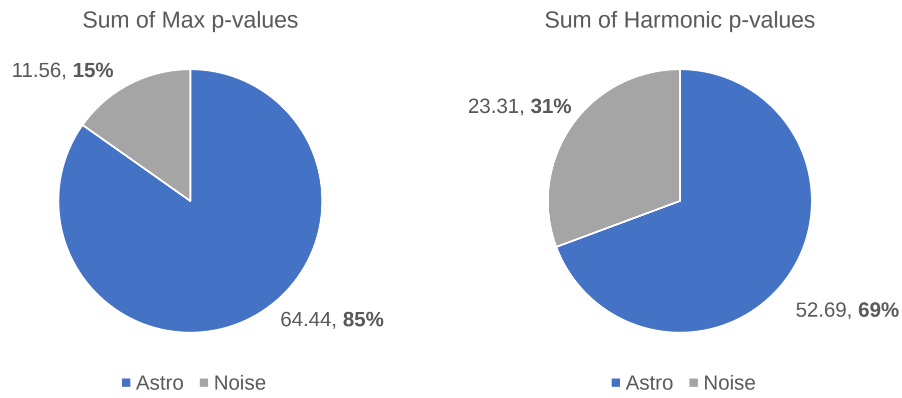

FIG. 5: Purity of 76 events detected by the Max method calculated from the sum of p_{astro} values from Max and Harmonic methods. The number estimated with the Harmonic method is much lower, which indicates that purity might be overestimated.

operative agreement PHY-1764464. Advanced LIGO was built under award PHY-0823459. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants PHY-0757058 and PHY-0823459. This work carries LIGO document number T2300203.

[1] LIGO Scientific Collaboration and Virgo Collaboration. GW150914: First results from the search for binary black hole coalescence with Advanced LIGO. *Phys. Rev. D*, 93:122003, Jun 2016.

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

[3] F. S. Broekgaarden. ChatGPT scores a bad birdie in counting gravitational-wave chirps. *arXiv:2303.17628*, Apr 2023.

[4] C. Capano, T. Dent, C. Hanna, et al. Systematic errors in estimation of gravitational-wave candidate significance. *Phys. Rev. D*, 96:082002, Oct 2017.

[5] S. A. Usman, A. H. Nitz, I. W. Harry, et al. The PyCBC search for gravitational waves from compact binary coalescence. *Classical and Quantum Gravity*, 33(21):215004, Oct 2016.

[6] S. Sachdev, S. Caudill, H. Fong, et al. The GstLAL Search Analysis Methods for Compact Binary Mergers in Advanced LIGO's Second and Advanced Virgo's First Observing Runs. *arXiv:1901.08580*, Jan 2019.

[7] F. Aubin, F. Brighenti, R. Chierici, et al. The MBTA pipeline for detecting compact binary coalescences in the third LIGO–Virgo Observing Run. *Classical and Quantum Gravity*, 38(9):095004, 2021.

[8] T. Venumadhav, B. Zackay, J. Roulet, et al. New search pipeline for compact binary mergers: Results for binary black holes in the first observing run of Advanced LIGO. *Phys. Rev. D*, 100:023011, Jul 2019.

[9] A. H. Nitz, S. Kumar, Y.-F. Wang, et al. 4-OGC: Catalog of gravitational waves from compact-binary mergers. *arXiv:2112.06878*, Dec 2021.

[10] S. Banagiri, C. P. L. Berry, G. S. Cabourn Davies, et al. A Unified p_{astro} for Gravitational Waves: Consistently Combining Information from Multiple Search Pipelines. *arXiv:2305.00071*, May 2023.

[11] LIGO Scientific Collaboration and Virgo Collaboration. GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs. *Phys. Rev. X*, 9:031040, Sep 2019.

[12] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[13] Daniel J. Wilson. The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200, 2019.

[14] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 1936.

[15] The LIGO Scientific Collaboration and the Virgo Collaboration. GWTC-2.1: Deep Extended Catalog of Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run. *arXiv:2108.01045*, May 2022.

[16] S. Olsen, T. Venumadhav, J. Mushkin, et al. New binary black hole mergers in the LIGO-Virgo O3a data. *Phys. Rev. D*, 106:043009, Aug 2022.

[17] W. M. Farr, J. R. Gair, I. Mandel, and C. Cutler. Counting and confusion: Bayesian rate estimation with multiple populations. *Phys. Rev. D*, 91:023005, Jan 2015.