Unified p_{astro} for gravitational waves: Consistently combining information from multiple search pipelines

Sharan Banagiri,[1,*] Christopher P. L. Berry,[2] Gareth S. Cabourn Davies,[3] Leo Tsukada,[4,5] and Zoheyr Doctor[1,6]

[1]*Center for Interdisciplinary Exploration and Research in Astrophysics, Northwestern University, 1800 Sherman Avenue, Evanston, Illinois 60201, USA*
[2]*SUPA, School of Physics and Astronomy, University of Glasgow, Glasgow G12 8QQ, United Kingdom*
[3]*University of Portsmouth, Portsmouth PO1 3FX, United Kingdom*
[4]*Department of Physics, The Pennsylvania State University, University Park, Pennsylvania State 16802, USA*
[5]*Institute for Gravitation and the Cosmos, The Pennsylvania State University, University Park, Pennsylvania State 16802, USA*
[6]*Department of Physics and Astronomy, Northwestern University, 2145 Sheridan Road, Evanston, Illinois 60201, USA*

(Received 28 April 2023; accepted 15 September 2023; published 31 October 2023; corrected 7 November 2023)

Recent gravitational-wave transient catalogs have used p_{astro}, the probability that a gravitational-wave candidate is astrophysical, to select interesting candidates for further analysis. Unlike false alarm rates, which exclusively capture the statistics of the instrumental noise triggers, p_{astro} incorporates the rate at which triggers are generated by both astrophysical signals and instrumental noise in estimating the probability that a candidate is astrophysical. Multiple search pipelines can independently calculate p_{astro}, each employing a specific data reduction. While the range of p_{astro} results can help indicate the range of uncertainties in its calculation, it complicates interpretation and subsequent analyses. We develop a statistical formalism to calculate a *unified* p_{astro} for gravitational-wave candidates, consistently accounting for triggers from all pipelines, thereby incorporating extra information about a signal that is not available with any one single pipeline. We demonstrate the properties of this method using a toy model and by application to the publicly available list of gravitational-wave candidates from the first half of the third LIGO-Virgo-KAGRA observing run. Adopting a unified p_{astro} for future catalogs would provide a simple and easy-to-interpret selection criterion that incorporates a more complete understanding of the strengths of the different search pipelines.

DOI: 10.1103/PhysRevD.108.083043

I. INTRODUCTION

The detection of gravitational waves (GWs) [1,2] by the Laser Interferometer Gravitational-Wave Observatory (LIGO) [3] and Virgo [4] detectors is the culmination of decades of research. Not only are sensitive detectors needed to the measure the GW signals [5,6], but sophisticated data analysis is needed to distinguish astrophysical signals from detector noise [7]. In the case of transient signals, such as those from compact binary coalescences (CBCs), detection algorithms identify candidate signals by matching template signals to the data [8] or by looking for coherent signals in multiple detectors [9]. Only after identifying GW candidates in detector data can we start to understand the astrophysical population of GW sources.

A fundamental question in any data analysis problem is the veracity of the signal or the effect it is considering. The statistical significance of a transient GW is generally assessed by calculating how likely the data would appear due to noise fluctuations. This may be quantified by the false alarm rate (FAR) or by a p value. Statistics like FAR are particularly well suited for making a first detection, where we do not know the population of signals. However, a more complete assessment of the probability that a candidate is real can be obtained by considering not only the FAR but also the *true* alarm rate (i.e., how often the algorithm would identify a real signal) whose calculation requires knowledge of the source population and our sensitivity to their signals, and hence is subject to additional uncertainties not inherent in the calculation of the FAR or p value. By combining the false and true alarm rates, we can calculate the probability of astrophysical origin p_{astro} for a candidate [10,11]. In the GW literature, p_{astro} was first used to estimate the astrophysical probability of GW150914 and the then-considered marginal trigger LVT151012 [12,13] (now known as GW151012 [14]). The probability of astrophysical origin p_{astro} directly addresses the key

*sharan.banagiri@northwestern.edu

question of how probable a candidate is to be real accounting for both our understanding of detector sensitivity and the population of GW sources.

We are now in the era of having many GW candidates, such that we are building an understanding of the source population [15]. GWs catalogs from both the LIGO-Virgo-KAGRA (LVK) Collaboration [2,14] and independent teams [16,17] now commonly use $p_{\rm astro}$ as a criterion to identify interesting candidates for further analysis. By the end of their third observing run (O3), the LVK Collaboration have reported 90 CBC candidates with $p_{\rm astro} > 0.5$ [2,14,18,19]. Despite the additional uncertainties inherent in its calculation, $p_{\rm astro}$ is important to consider now that we now have observations from multiple observing runs of different sensitivity [6] and observations of different types of sources [1,2,20,21]. At a given FAR, a candidate from the most recent O3 run [22,23] is more probable to be real than a candidate from the (less sensitive) first observing run [24]. Similarly, at a given FAR, a candidate consistent with originating from a $30M_\odot + 30M_\odot$ binary black hole (BBH) merger is more probable to be real than a candidate consistent with a binary neutron star (BNS) origin [25] because the former can be detected from a greater distance and so are more frequently observed [2]. Since $p_{\rm astro}$ accounts for the true alarm rate, it can naturally be used when compiling a heterogeneous catalog of diverse sources from different observing runs.

During O3, the LVK Collaboration used four search pipelines for its final analysis [2], with each calculating its own $p_{\rm astro}$: there are three CBC matched-filtered pipelines (GstLAL [25–29], PyCBC [30–35] and MBTA [36–38]), and non-template-based search pipeline (cWB [39–41]) that makes minimal assumptions about the transient signals. Multiple analyses with different choices help to explore the GW-signal space more thoroughly, making it easier to find a diverse range of signals. Unfortunately, this complicates the use of $p_{\rm astro}$ when interpreting current GW candidates in three ways. First, there are differences in the assumptions about the source population between analyses that mean that results are not directly comparable. Second, as illustrated in Fig. 6 of GWTC-3 [2], pipelines can have different relative sensitivities, which may provide valuable insights into whether the candidate is real. This information is not incorporated in individual $p_{\rm astro}$ estimates. Third, it means there is no single assessment of the significance of a candidate, complicating the calculation of contamination and assessment of search sensitivity (and hence selection effects, which are essential for population studies). Indeed, as can be seen in Table 1 of GWTC-3 [2], pipelines can often find a different set of triggers over the same stretch of data and can sometimes calculate significantly different $p_{\rm astro}$ values for the same candidate. To fully make use of the different search pipelines, it is necessary to combine their results to produce a single $p_{\rm astro}$ value for each candidate.

Some early work in combining results from multiple pipelines was done in the predetection era. For example, developing a way to estimate generalized frequentist upper limits with multiple pipelines [42], and developing unified detection statistics [43]. In this paper, we present a framework that can combine the results from multiple detection algorithms to produce a unified $p_{\rm astro}$ employing information about the correlation between pipelines. This means that catalogs can be produced consistently using a simple threshold on this statistic and that downstream users may more straightforwardly calculate the contamination fraction for such a catalog. This statistic is amenable to use in population inferences that incorporate low-significance candidates [44,45]. Crucially, as the formalism accounts for the correlations between different pipeline outputs (or lack thereof), we can make use of the full ensemble of results when evaluating whether a candidate is real. We present an illustration of the properties of our $p_{\rm astro}$ formalism, and provide a proof-of-concept example of its use with real GW data from the first half of O3 (henceforth called O3a) analyzed for GWTC-2.1 [19].

In Sec. II we define $p_{\rm astro}$ from first principles and describe its connection with the Poisson mixture-model formalism, first developed by Farr *et al.* [10] (henceforth called the FGMC method) in the GW context. Section III extends this formalism to calculate unified $p_{\rm astro}$ from multiple search pipelines. We demonstrate some of its useful properties by the means of a simple toy model in Sec. IV, followed by an illustrative application to triggers from GWTC-2.1 in Sec. V. This illustration highlights the tools that need to be developed in order to produce a reliable unified $p_{\rm astro}$ for use in future GW catalogs. We discuss the applications and extensions of our formalism in Sec. VI.

II. DEFINING $p_{\rm astro}$

In this section, we present a pedagogical derivation for $p_{\rm astro}$ as commonly seen in literature [10,12,13,25], connecting it to the Poisson mixture-model FGMC formalism. In Sec. III, we explain how to extend this for the case of multiple search pipelines.

The primary goal of a search pipeline is to identify candidates of astrophysical origin. Pipelines usually accomplish this by maximizing some statistics over a small stretch of data, and identifying a particular candidate, which we will call a trigger. We assume that the stretch of data contains at most one signal. Every trigger has a detection statistic x associated with it by the search pipeline.

The distribution of the detection statistic under pure noise lets us calculate the FAR distribution of the pipeline. In the context of GW analysis, this can be done by the bootstrapping method of time slides, where the data stream from one detector is shifted with respect to another by more

than the light-travel time between them [46], or by constructing a model of the noise using triggers that are not coincident between detectors [47].

We define $p_{\rm astro}$ as the probability that a particular trigger is caused by an astrophysical signal, as opposed to a noise fluctuation or terrestrial contamination. Hence $p_{\rm astro}$ is properly associated with the trigger and its detection statistic (one specific reduction of the data) rather than directly the underlying data itself. When different pipelines analyze the data, they make different analysis choices which give them differing sensitivities, and so it is possible that they yield different triggers and $p_{\rm astro}$ values.

Since a trigger can be caused by an astrophysical signal or by noise, $p_{\rm astro}$ depends on the posterior probability for these two hypotheses. We define $p(\mathcal{S}|x,\Phi_{\rm s},\Phi_{\rm n})$ as the posterior probability for the signal hypothesis $\mathcal{S}$, and $p(\varnothing|x,\Phi_{\rm n})$ as the posterior probability for the noise hypothesis $\varnothing$. Both these probabilities are conditioned on the detection statistic of the trigger x, and assume some signal and noise parameters, $\Phi_{\rm s}$ and $\Phi_{\rm n}$. We can then write $p_{\rm astro}$ for the trigger as the normalized probability that it is astrophysical,

$$p_{\rm astro}(x) = \frac{p(\mathcal{S}|x,\Phi_{\rm s},\Phi_{\rm n})}{p(\mathcal{S}|x,\Phi_{\rm s},\Phi_{\rm n}) + p(\varnothing|x,\Phi_{\rm n})}. \tag{1}$$

We shall suppress the signal and noise parameters from here on for the purpose of clarity. A value $p_{\rm astro}=1$ implies perfect confidence about the presence of a signal in an underlying segment of data, while $p_{\rm astro}=0$ implies perfect confidence of its absence, and $p_{\rm astro}=0.5$ implies no preference between the signal and noise hypothesis.

The posterior probabilities for two cases are difficult to compute directly. What is more easily accessible is the likelihood of the trigger x, under the signal and noise hypotheses. Therefore using Bayes' theorem, we write $p(\mathcal{S}|x)$ and $p(\varnothing|x)$ as

$$\begin{aligned} p(\mathcal{S}|x) &= \frac{\pi_{\rm s} p(x|\mathcal{S})}{Z(d)}, \\ p(\varnothing|x) &= \frac{\pi_{\rm n} p(x|\varnothing)}{Z(d)}, \end{aligned} \tag{2}$$

where $p(x|\mathcal{S})$ and $p(x|\varnothing)$ are the likelihoods for getting a trigger with statistic x under the signal and noise hypothesis, respectively; $\pi_{\rm s}$ and $\pi_{\rm n}$ are the corresponding priors, and $Z(d)$ is the Bayesian evidence. We can then write $p_{\rm astro}$ as

$$p_{\rm astro}(x) = \frac{\pi_{\rm s} p(x|\mathcal{S})}{\pi_{\rm s} p(x|\mathcal{S}) + \pi_{\rm n} p(x|\varnothing)}. \tag{3}$$

Suppose that we analyze a long stretch of data with a single search pipeline. Let $R_{\rm s}$ and $R_{\rm n}$ be the rates of astrophysical and noise-only triggers, respectively, assuming a given detector sensitivity, operating characteristics, and algorithmic choice (including any preliminary thresholds that are applied) made by the search pipeline. We define the signal and noise count parameters, $\Lambda_{\rm s}$ and $\Lambda_{\rm n}$, respectively. These are the mean number of astrophysical and noise triggers (above the predefined thresholds) for an observing duration T (not necessarily the number of triggers in any particular realization of the data) such that

$$\Lambda_{\rm s} = R_{\rm s} T \quad \text{and} \quad \Lambda_{\rm n} = R_{\rm n} T. \tag{4}$$

The total number of triggers N will follow a Poisson distribution with a count parameter $\Lambda = \Lambda_{\rm s} + \Lambda_{\rm n}$:

$$p(N|\Lambda_{\rm s},\Lambda_{\rm n}) = \frac{(\Lambda_{\rm s}+\Lambda_{\rm n})^N e^{-(\Lambda_{\rm s}+\Lambda_{\rm n})}}{N!}. \tag{5}$$

For computational reasons, it is common to require that the trigger has to pass some primary statistical threshold, before being used for further analysis. For example, the GWTC-2.1 [19] and GWTC-3 [2] analyses considered only triggers that had a FAR ≤ 2 day^{-1}. Our formalism does not make any strong assumptions about the preliminary threshold, as long as it is consistently used among all pipelines. Thresholds are incorporated by suitably defining the counts $\Lambda_{\rm s}$ and $\Lambda_{\rm s}$ as for triggers that pass the threshold.

The count parameters are generally assumed to be unknown and have to be measured empirically using the triggers. To do this we use the Poisson mixture-model FGMC formalism:

$$p(\{x\}|\Lambda_{\rm s},\Lambda_{\rm n},N) = \prod_i^N [\pi_{\rm s} p(x_i|\mathcal{S}) + \pi_{\rm n} p(x_i|\varnothing)], \tag{6}$$

where $\{x\}$ denotes the set of all triggers found in the data. Conditioned upon $\Lambda_{\rm s}$ and $\Lambda_{\rm n}$, the prior probabilities are just the relative rate of the triggers of that category,

$$\pi_{\rm s} = \frac{\Lambda_{\rm s}}{\Lambda_{\rm s}+\Lambda_{\rm n}}, \qquad \pi_{\rm n} = \frac{\Lambda_{\rm n}}{\Lambda_{\rm s}+\Lambda_{\rm n}}. \tag{7}$$

Therefore, Eq. (6) becomes

$$p(\{x\}|\Lambda_{\rm s},\Lambda_{\rm n},N) = \frac{\prod_i^N [\Lambda_{\rm s} p(x_i|\mathcal{S}) + \Lambda_{\rm n} p(x_i|\varnothing)]}{(\Lambda_{\rm s}+\Lambda_{\rm n})^N}. \tag{8}$$

Using Bayes' theorem, we can relate the posterior distribution on the count parameters with the Poisson likelihood,

$$\begin{aligned} p(\Lambda_{\rm s},\Lambda_{\rm n}|\{x\},N) &\propto p(\{x\},N|\Lambda_{\rm s},\Lambda_{\rm n})\pi(\Lambda_{\rm s},\Lambda_{\rm n}) \\ &\propto p(\{x\}|N,\Lambda_{\rm s},\Lambda_{\rm n}) p(N|\Lambda_{\rm s},\Lambda_{\rm n})\pi(\Lambda_{\rm s},\Lambda_{\rm n}), \end{aligned} \tag{9}$$

where $\pi(\Lambda_{\rm s},\Lambda_{\rm n})$ is the prior on the mean count. Putting this together with Eq. (5), we obtain the posterior distributions for $\Lambda_{\rm s}$ and $\Lambda_{\rm n}$,

$$p(\Lambda_s,\Lambda_n|\{x\},N) \propto e^{-(\Lambda_n+\Lambda_s)}\pi(\Lambda_s,\Lambda_n) \times \prod_i^N \{\Lambda_s p(x_i|\mathcal{S}) + \Lambda_n p(x_i|\varnothing)\}. \quad (10)$$

Finally, combining Eqs. (3) and (7), we can calculate a $p_{\rm astro}$ that marginalizes over the posterior of Λ_s and Λ_n,

$$p_{\rm astro}(x) = \int d\Lambda_s d\Lambda_n \frac{\Lambda_s p(x|\mathcal{S}) p(\Lambda_s,\Lambda_n|\{x\},N)}{\Lambda_s p(x|\mathcal{S}) + \Lambda_n p(x|\varnothing)}. \quad (11)$$

This fundamental framework has been used by all search pipelines both by the LVK Collaboration and other groups in compiling GW catalogs [2,12,13,16,17,19,48–50].

III. TOWARDS A UNIFIED $p_{\rm astro}$

We now develop a way to combine information from multiple pipelines for calculating a unified $p_{\rm astro}$. Suppose that we have several search pipelines that yield triggers after running over the same underlying data. Consider any two corresponding triggers x^α and x^β from pipelines α and β; as these triggers are produced by different pipelines they may be associated with different statistics, but these statistics will be correlated depending upon the relative sensitivities of the two pipelines. Understanding the correlations between pipelines is key to understanding how to construct a unified $p_{\rm astro}$.

The correlations between different search pipelines are not typically accounted for when compiling GW results. In some analyses, the FAR is multiplied by a trials factor equal to the number of different searches [51,52]. However, this is a conservative choice: it would be correct if results were noise triggers that were uncorrelated, but in general, we would expect some correlation since search pipelines are searching for similar signals in the data. This correlation should reduce the effective trials factor (in the limit of running two identical pipelines there would be no need to add a trials factor). Our framework accounts for correlations in both noise triggers and signal triggers to construct a unified $p_{\rm astro}$.

Let us define $\vec{x} = \{x^\alpha, x^\beta, ...\}$ as the triggers that correspond with each other from a series of different pipelines. For the rest of this paper, a vector usually means a vector of pipeline outputs, such that $\{\vec{x}\}$ is the set of all triggers from all the pipelines in the stretch of data being analyzed, and a latin index is used to indicate the individual data segment or trigger. As in Eq. (2), the probability $p(\mathcal{S}|\vec{x})$ is given by

$$p(\mathcal{S}|\vec{x}) \propto \Lambda_s p(\vec{x}|\mathcal{S}). \quad (12)$$

Similarly, for the noise hypothesis,

$$p(\varnothing|\vec{x}) \propto \Lambda_n p(\vec{x}|\varnothing). \quad (13)$$

Here, $p(\vec{x}|\mathcal{S})$ and $p(\vec{x}|\varnothing)$ are joint likelihoods for obtaining $\vec{x} = \{x^\alpha, x^\beta, ...\}$ and are dependent on the aforementioned correlations between trigger statistics across pipelines. If we learn these joint likelihood distributions, using simulated signals and noise triggers, we can calculate a unified $p_{\rm astro}$.

While several methods to learn arbitrary distributions exist, we will primarily use kernel density estimation (KDE) methods from SCIKIT-LEARN [53] to learn $p(\vec{x}|\mathcal{S})$ and $p(\vec{x}|\varnothing)$. The choice of method used to reconstruct the distribution is important, and the distributions must be properly characterized to successfully calculate a reliable $p_{\rm astro}$. However, for the purposes of illustrating the framework needed to compute a unified $p_{\rm astro}$, this method may be treated as a black box.

Once the distribution of these correlations is learned, one can modify Eq. (10) for a joint FGMC estimate

$$p(\Lambda_s,\Lambda_n|\{\vec{x}\},N) \propto e^{-(\Lambda_n+\Lambda_s)}\pi(\Lambda_s,\Lambda_n) \times \prod_i^N \{\Lambda_s p(\vec{x}_i|\mathcal{S}) + \Lambda_n p(\vec{x}_i|\varnothing)\}. \quad (14)$$

Equation (11) can then be modified to calculate a unified $p_{\rm astro}$ marginalized over signal and noise counts,

$$p_{\rm astro}(\vec{x}) = \int d\Lambda_s d\Lambda_n \frac{\Lambda_s p(\vec{x}|\mathcal{S}) p(\Lambda_s,\Lambda_n|\{\vec{x}\},N)}{\Lambda_s p(\vec{x}|\mathcal{S}) + \Lambda_n p(\vec{x}|\varnothing)}. \quad (15)$$

This equation therefore incorporates correlations between pipelines through the joint likelihoods $p(\vec{x}|\mathcal{S})$ and $p(\vec{x}|\varnothing)$.

The role of correlations in calculating $p_{\rm astro}$ may be understood by considering a few idealized cases. If we had two pipelines that looked for the same type of signal, but had different sensitivities to noise triggers, we might expect to be more certain in a candidate (assign a higher $p_{\rm astro}$) if there are corresponding triggers from both pipelines. However, if a candidate was identified by one pipeline, and not by another which has greater sensitivity to that type of signal, we might expect to suspect it as a false alarm (assign a lower $p_{\rm astro}$). In the following subsections, we will demonstrate the properties of our unified $p_{\rm astro}$ in some illustrative cases.

IV. TOY MODEL

In this section we illustrate and test the unified $p_{\rm astro}$ method using a simple toy model. In Sec. V, we will then apply the method on real GW data, albeit in a simplified analysis.

The toy model data is generated in the form of segments, each of which consists of four data points with values drawn from a standard normal distribution. A segment might also contain a signal with a probability proportional to Λ_s. Signals are simulated by adding a value λ_s to one of the data points in the segment. We assume that λ_s is known.

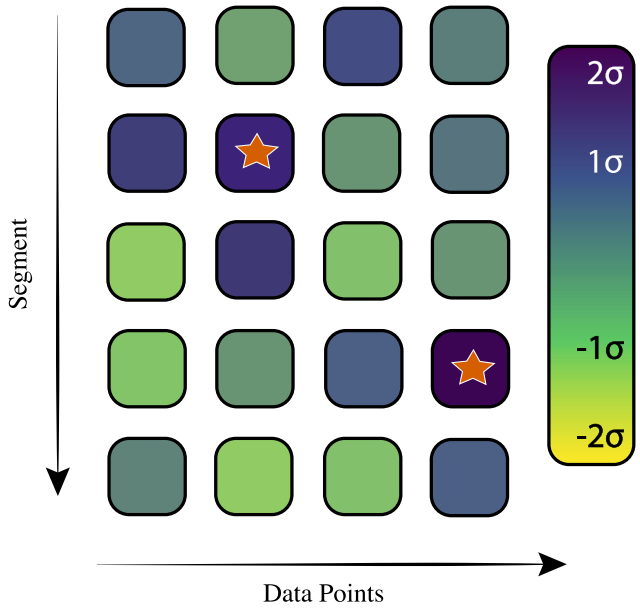


FIG. 1. A schematic of the toy model data. Each row is a segment, consisting of four data points indicated by the rounded rectangles. The noise in the segments is drawn from a standard normal distribution, with its value indicated by the color scale of the rounded rectangle. Segments 2 and 4 from the top in this example contain a signal, added randomly to one of the four data points indicated by the vermilion star.

The goal of an FGMC-type analysis is to estimate the fraction Λ_s. Figure 1 demonstrates a schematic of the data generated in the toy model.

We construct four simple pipelines to analyze this data on a segment-by-segment basis. The pipelines are all based on a maximum-likelihood estimate assuming Gaussian noise statistics, and the pipelines all assume that the noise is drawn from a standard normal with $\sigma = 1$. For each segment, the pipelines calculate the noise likelihood

$$\mathcal{L}_n \propto \exp\left[-\frac{1}{2}\frac{(\sum_{i=1}^{k} x_i)^2}{k\sigma^2}\right], \tag{16}$$

and the signal likelihood, assuming the signal value λ_s is known

$$\mathcal{L}_s \propto \exp\left[-\frac{1}{2}\frac{(\sum_{i=1}^{k} x_i - \lambda_s)^2}{k\sigma^2}\right], \tag{17}$$

where x_i are the analyzed data points in the segment. The pipelines might use different number of data points, corresponding to the number $k \leq 4$. Each pipeline calculates a p value for every segment (under the noise hypothesis) which is the main input for the unified p_{astro} analysis.

The four pipelines are

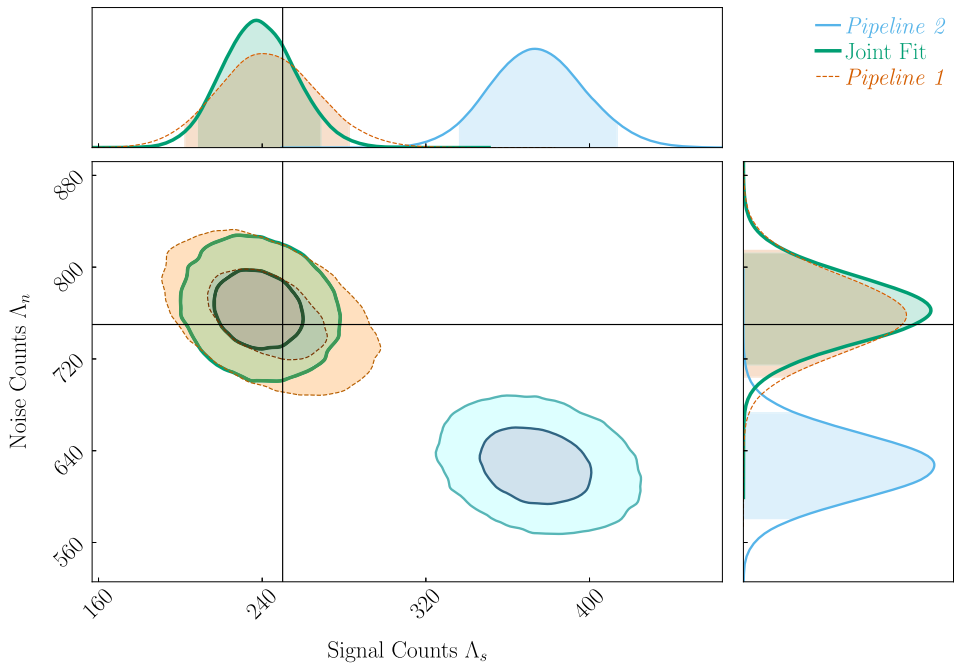
(1) *Pipeline 1*: This pipeline makes no assumption on which of the four data points in a segment has the signal, effectively marginalizing over the position of the signal. Therefore, $k = 4$ in Eqs. (16) and (17).
(2) *Pipeline 2*: This pipeline still assumes Gaussian statistics but only considers the data point with the loudest value of the four and calculates its likelihood. The is equivalent to setting $k = 1$ in Eqs. (16) and (17), but replacing x_i with $\max\{x_i\}$. This pipeline will be more effective at detecting a loud signal but will generally overestimate the number of signals.
(3) *Pipeline 3*: This only uses the first two data points of a segment. Therefore, $k = 2$ in Eqs. (16) and (17). Since in our toy model, we know that each data point is equally probable to contain a signal, we expect that this pipeline will witness only about half of the total signals, and consequently its count parameter will be around half the true value. Thereby, this is analogous to a search pipeline that will be sensitive to only a subset of GW signals (e.g., those from high-mass systems).
(4) *Pipeline 4*: Similar to pipeline 3 except it only uses the last two data points. Therefore pipeline 4 is statistically independent from pipeline 3 while still seeing only half of the total signals. Combining this pipeline with pipeline 3 is the simplest example of a correlation between pipelines.

For each pipeline, we also estimate the single-pipeline signal and noise counts using the signal and noise likelihoods.

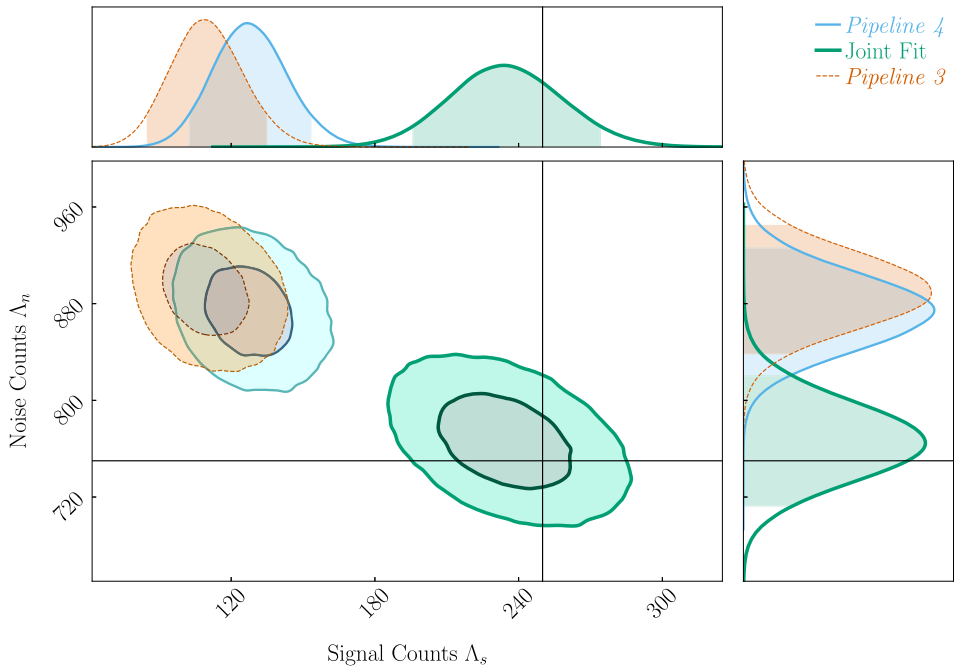
Using the p value from the pipelines as their detection statistics, we will calculate the joint likelihoods using a KDE as described in Sec. III. To do this we create a number of segments containing a signal to train a signal KDE, and similarly train a noise KDE using segments that do not contain a signal. Then for a simulation where Λ_s is unknown, we can score the outputs of the pipelines against the KDEs to perform a unified FGMC analysis using Eq. (14), and measure the noise and signal counts, Λ_n and Λ_s.

The algorithmic choices made by the pipelines mean that their results differ from each other, and can introduce a bias in their estimate of signal counts (and thereby p_{astro}). For instance, consider pipeline 3, which can only see about half of the total signals. When combining the output of the pipelines, the KDEs do not know *a priori* about these biases and will have to learn it from the triggers. In a joint analysis between pipeline 3 and pipeline 4, the signal KDE can learn that only about half the analyzed segments will have a low p value in pipeline 3 (i.e., when a signal is added to the first two data points) and that these are the segments in which pipeline 4 will likely give a high p value (and vice versa).

Figure 2 shows an FGMC analysis for two separate toy model analyses. In both Figs. 2(a) and 2(b), we separately simulated 1000 segments of data with about a quarter ($\Lambda_s = 250$) of the segments having a signal with $\lambda_s = 3$. Figure 2(a) shows the signal and noise counts estimated by



(a)



(b)

FIG. 2. Toy-model exploration of the unified p_{astro} formalism. A signal value $\lambda_s = 3$ was used in both cases. In both figures, the solid vertical and horizontal lines indicate the real number of segments with (and without) a signal. The shaded regions in the marginal plots are 90% uncertainty levels while the two-dimensional contours correspond to 50% and 90% levels. (a) Analysis with pipeline 1 and pipeline 2. The horizontal and vertical axes are the measured signal and noise count parameters. Pipeline 2 is clearly biased as it only uses the loudest data point of a segment for analysis, but the unified FGMC analysis performing a joint fit can correct for it. (b) Analysis with pipeline 3 and pipeline 4. While both pipelines only see half the simulated signals, the joint analysis learns to correct for this from the correlations in the joint signal distributions.

pipeline 1 and pipeline 2, along with a unified fit. As expected, pipeline 2 shows a bias in its recovery which nevertheless the joint analysis can correct for. Figure 2(b) shows the results of a toy-model analysis with pipeline 3 and pipeline 4. Both pipelines only see about half of the simulated signals, by design. However, the joint analysis learns this bias through the signal KDE and can account for it.

This section shows, using a simple example, how outputs from multiple pipelines can be combined using the unified formalism that we have developed. It also shows how certain kinds of biases can be corrected when multiple pipelines are joined together, by drawing information from the correlations (or lack thereof) between the outputs of the pipelines. These examples demonstrate the robustness of the unified $p_{\rm astro}$ method.

V. APPLICATION TO GWTC-2.1

We now test the unified $p_{\rm astro}$ method with real GW triggers (henceforth *on-source* triggers) by applying this to O3a results from GWTC-2.1 [21]. GWTC-2.1 applies a preliminary cut of FAR ≤ 2 day^{-1}, which we also adopt. While this analysis uses real data, it is only intended to be illustrative, and we adopt some simplifications for computational simplicity.

We use results from the GstLAL [26,27,29] and PyCBC [32–34] pipelines, and restrict ourselves to triggers that are at least Hanford-Livingston coincident (i.e., present in data from LIGO Hanford and Livingston at a minimum). GWTC-2.1 [21] uses two versions of the PyCBC pipelines; referred to as PyCBC-broad and PyCBC-BBH; in this analysis we only use the former version, which we will simply refer to as PyCBC.[1] While our formalism is general, we focus solely on CBC signals; a choice that is primarily driven by prudence, as we have only seen the population of CBC signals so far [2]. Furthermore, for the sake of simplicity we will only use BBH signals for the training; this will also serve as a test to see if the method can distinguish sufficiently between signal types and to test overfitting. These simplifications make it easier to illustrate the behavior of the unified $p_{\rm astro}$, and highlight the key considerations for an analysis to be performed for future GW catalogs.

We use pipeline settings that are as similar as possible to the runs in GWTC-2.1 [21]. A crucial factor to consider is that the search pipelines have to be run in a way that we can establish a one-to-one correspondence between their triggers so that we can learn the correlations between them. This is necessary even in the case where one pipeline registers a trigger and the other does not because the absence of the trigger is itself useful information. While one-to-one correspondence is straightforward for simulated signal triggers (described in Sec. V A), it can be somewhat ambiguous to define in the case of noise triggers (implementation details are explained in Sec. V B).

In order to calculate a unified $p_{\rm astro}$, Eqs. (14) and (15) require us to construct the joint distribution of a statistic from each pipeline. We choose to use the FAR as the base statistic here as it is easily interpretable, commonly used by all pipelines and informative of the relative significance of triggers. However, the dynamic range of FARs can be large, and we therefore define a statistic β that is related to the logarithm of the inverse FAR,

$$\beta(\mathrm{FAR}) = \log\left(1 + \frac{\kappa}{\mathrm{FAR}}\right), \tag{18}$$

where κ is a scaling constant, which we set to a value of $\kappa = 100$ yr^{-1}. This number is set to give a good dynamic range to β but our results are insensitive to its actual value. Higher values of β correspond to more significant candidates.

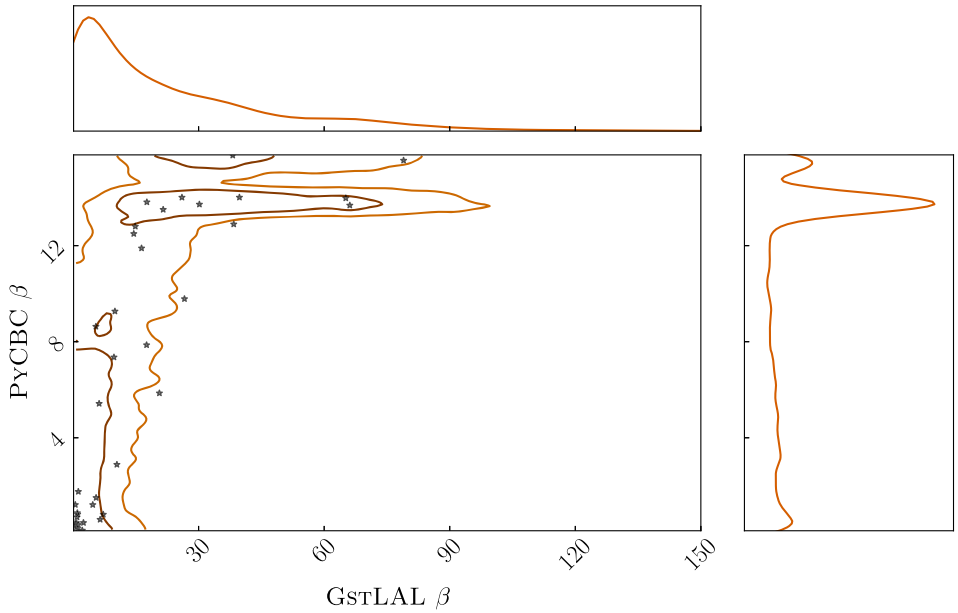
We construct three cases each for the noise and signal hypotheses, corresponding to triggers that are registered (i) only by GstLAL, (ii) only by PyCBC, and (iii) by both pipelines. We find that separate modeling the three cases is necessary in order to achieve a faithful fit of signal triggers, which can cover a wide range on the β space and show significant structure. In addition, this allows us to consider the case where one pipeline would not see a signal but the other does. For example, in the GWTC-2.1 analysis [19], PyCBC did not assign FARs to events which triggered in a single detector (although this is now possible [54]), but GstLAL did. In the future, a more sophisticated fitting method like a convolutional neural network or a random forest classifier might make multiple separate fits unnecessary.

The distributions are usually normalized such that the total probability over its parameter space is one. Therefore, when considering multiple separate cases we need to renormalize them to account for the relative probability of the class they represent. We do this by multiplying the output of the distribution with the fraction of signal or noise triggers that fall in that particular class. We find that getting a joint trigger is ≈104 times more likely under the signal hypothesis than the noise hypothesis. Meanwhile, a GstLAL-only (PyCBC-only) trigger is ≈2.77 (≈3.87) times more likely under the noise hypothesis than the signal hypothesis. The numbers quoted correspond to triggers that pass a 2 day^{-1} FAR threshold.

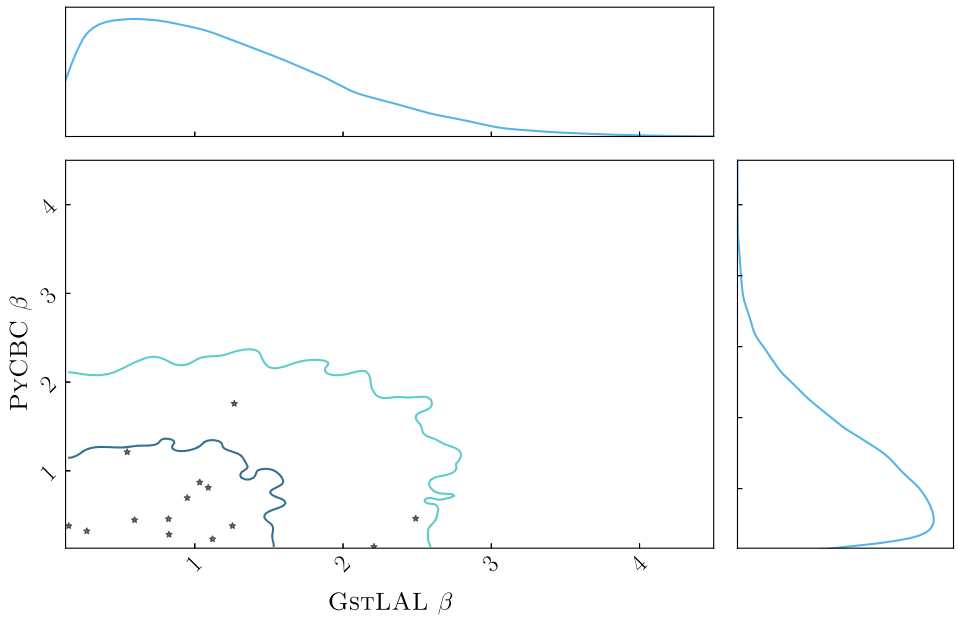
A. Signal distribution

The signal distributed is generated by fitting simulated software signals or *injections* that are added to the detector noise, to a KDE. We use the same simulated signals as from the GWTC-3 analysis [55], whose distribution is described in detail in Appendix C 7 (the injection row of Table X) of

[1]Our approach could be used to combine the PyCBC-broad and PyCBC-BBH results too; we pick the GstLAL and PyCBC-broad results to illustrate our method to show how it can be used across different pipelines.



(a)



(b)

FIG. 3. Joint likelihood distributions for the signal and noise hypothesis. The spans of the signal and noise distributions span are dissimilar. The signal distribution extends to much larger β values than the noise distribution. The contours correspond to 50% and 90% levels in both plots. (a) The signal distribution $p(\vec{x}_i|\mathcal{S})$ learned using a KDE to the simulated signals that are commonly found by both GstLAL and PyCBC. The stars are the O3a on-source triggers from GWTC-2.1 found by both pipelines [19]. The upper cutoff of the PyCBC distribution that is visible in the plot comes from the fact that the pipeline places a limit on the FAR based on the number of time slides performed. This translates to a cutoff at $\beta \approx 15.76$ and results in a bump in the PyCBC distribution at these values of β. (b) The noise distribution $p(\vec{x}_i|\varnothing)$ fit to a 2-dimensional Gaussian to the common noise triggers. The stars are the O3a on-source triggers from GWTC-2.1 found by both pipelines [19].

the GWTC-3 paper [2]. In particular, the source masses are distributed as

$$p(m_1) \propto m_1^{-2.35},$$
$$p(m_2|m_1) \propto m_2. \tag{19}$$

The redshift distribution is flat in comoving volume, with a maximum redshift of 1.9, assuming a Λ cold dark matter cosmology [56]. We restrict ourselves to injections with source-frame component masses $3M_\odot$–$100M_\odot$ to focus solely on BBHs. The simulated population is not a close match to the inferred astrophysical population [15,57]. However, it shares broad characteristics and is simple to use. Therefore it suffices for our illustrative calculation. Using a population model that more closely resembles the underlying population should enable more accurate calculation of the true alarm rate, and hence $p_{\rm astro}$.

In any KDE-based analysis, the choice of the KDE bandwidth is important, especially for multidimensional distributions with intricate shapes. In our illustrative analysis, the bandwidth was picked manually by checking that the KDE approximates well the distribution of injections. We verified that the one-dimensional cumulative distributions of the KDE and the injection set match, and also cross-checked the KDE against a KDE using the Silverman rule of thumb [58]. Ultimately, a realistic application would need a more sophisticated fitting scheme, for example, KDEs fit using an iterative approach 59]] or a machine-learning approach such as using normalizing flows.

The top plot of Fig. 3(a) shows the joint distribution of signals that are found by both PyCBC and GstLAL, and Fig. 4 shows the distribution of signal triggers that are found only by one pipeline.

B. Noise distribution

The noise distribution is fit using data with a nonphysical time shift higher than the light travel time between detectors to remove correlations between any real signals. Generating this data is computationally expensive, therefore, in this proof-of-concept analysis, we use O3a data with a single time shift where LIGO Livingston and Virgo data streams were shifted by 0.62831 and 0.31415 s, respectively, in relation to LIGO Hanford.

To define corresponding noise triggers between the pipelines, we match the times of the time-shifted triggers

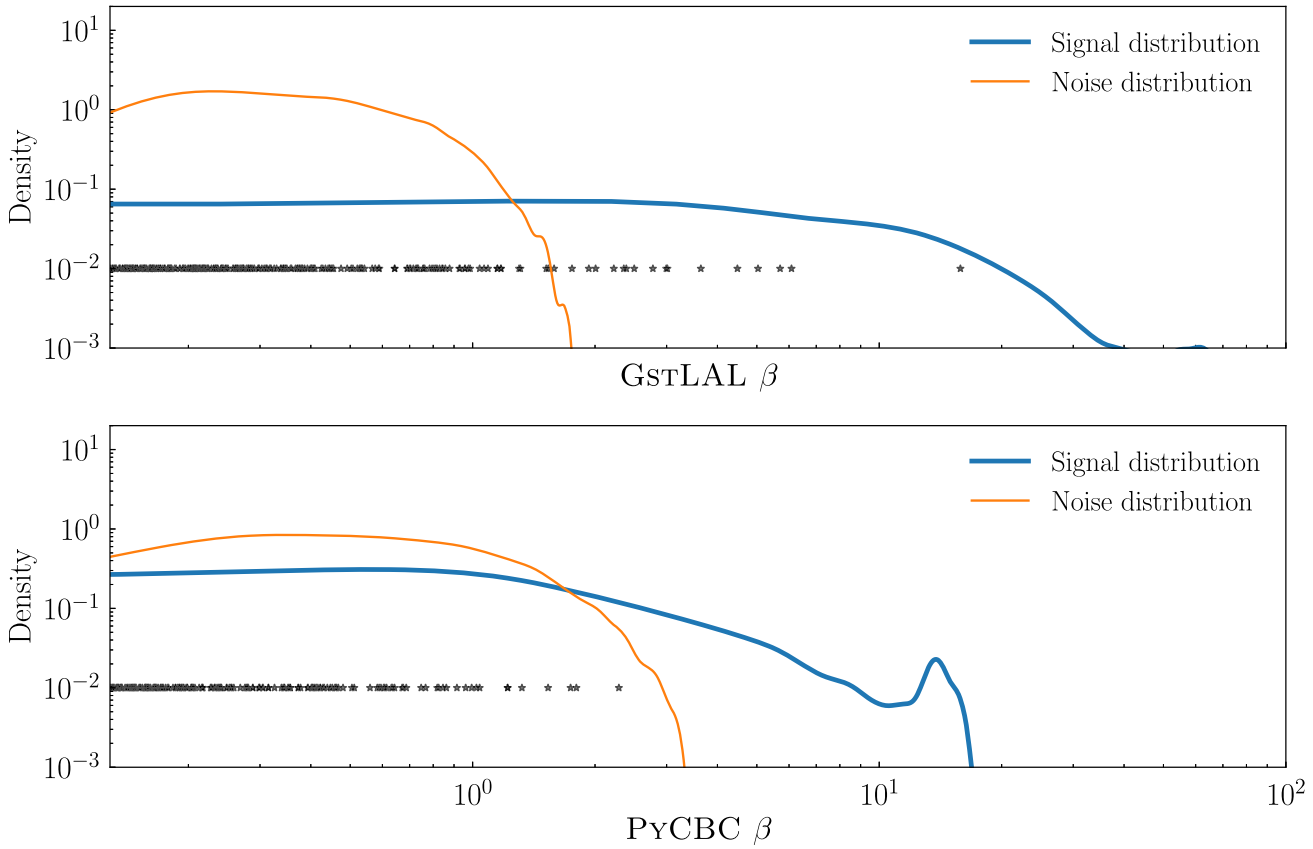


FIG. 4. The distribution of triggers that were found by one pipeline only. The top plot shows the noise and signal distribution of GstLAL, while the bottom plot shows the PyCBC distribution of triggers. The black stars are again the on-source O3a triggers from GWTC-2.1 that have only been found by the corresponding pipeline [19].

within a window of 1 s, comparable to the typical window used for grouping triggers in O3 [2]. This procedure gives us 256 GstLAL triggers and 348 PyCBC triggers that pass the FAR threshold of 2 day^{-1}; of these, only 12 triggers are common between pipelines. For a unified $p_{\rm astro}$ analysis intended to be used in production of a GW catalog, a larger set of noise triggers would be required.

The small number of noise triggers common between the pipelines makes it difficult to accurately reconstruct the distribution. However, as our primary goal is to illustrate unified $p_{\rm astro}$ method, rather than to produce a full set of results, these triggers are sufficient for completing a demonstrative calculation. To make use of the small number of triggers that are common between the pipelines, we adopt a bootstrapping procedure to generate more triggers. To do this, we first fit triggers that pass a 10 day^{-1} cut to a KDE, and draw 10^4 triggers that pass a 2 day^{-1} cut. This corresponds to doing ≈17 time slides in total. To avoid overfitting the noise triggers, we choose to employ a truncated Gaussian fit to them, instead of a KDE. The assumption of a Gaussian is purely for simplicity, and more careful choice of estimating the noise distribution would be needed for a proper analysis. The single-pipeline noise triggers are fit using a truncated Gaussian with mean $\mu = \beta(2\ \text{day}^{-1})$ and a standard deviation calculated from the noise triggers. The joint noise distribution is fit with a 2-dimensional truncated Gaussian with $\mu = \beta(2\ \text{day}^{-1})$ and the covariance matrix calculated from the noise triggers. Figure 3(b) shows the distribution of joint triggers, while Fig. 4 shows the single-pipeline triggers in this bootstrapped data set.

The sparseness of noise triggers means that the noise distribution is not accurately reconstructed. This is liable to give biased values for $p_{\rm astro}$: where the noise likelihood is underestimated, $p_{\rm astro}$ will be underestimated, and where the noise likelihood is overestimated, $p_{\rm astro}$ will be underestimated. However, the goal of this application is to show that the unified $p_{\rm astro}$ formalism yields sensible results to illustrate what would be needed for future analyses. A realistic application of this process to GW data will likely require multiple time shifts or some other process to generate a larger number of high-fidelity noise triggers, and then a careful reconstruction of the shape of the noise distribution.

C. Computing a unified $p_{\rm astro}$

After restricting ourselves to ones that are at least Hanford-Livingston coincident, and have been found be at least one of GstLAL or PyCBC, we are left with 553 on-source triggers from the GWTC-2.1 O3a results [2].

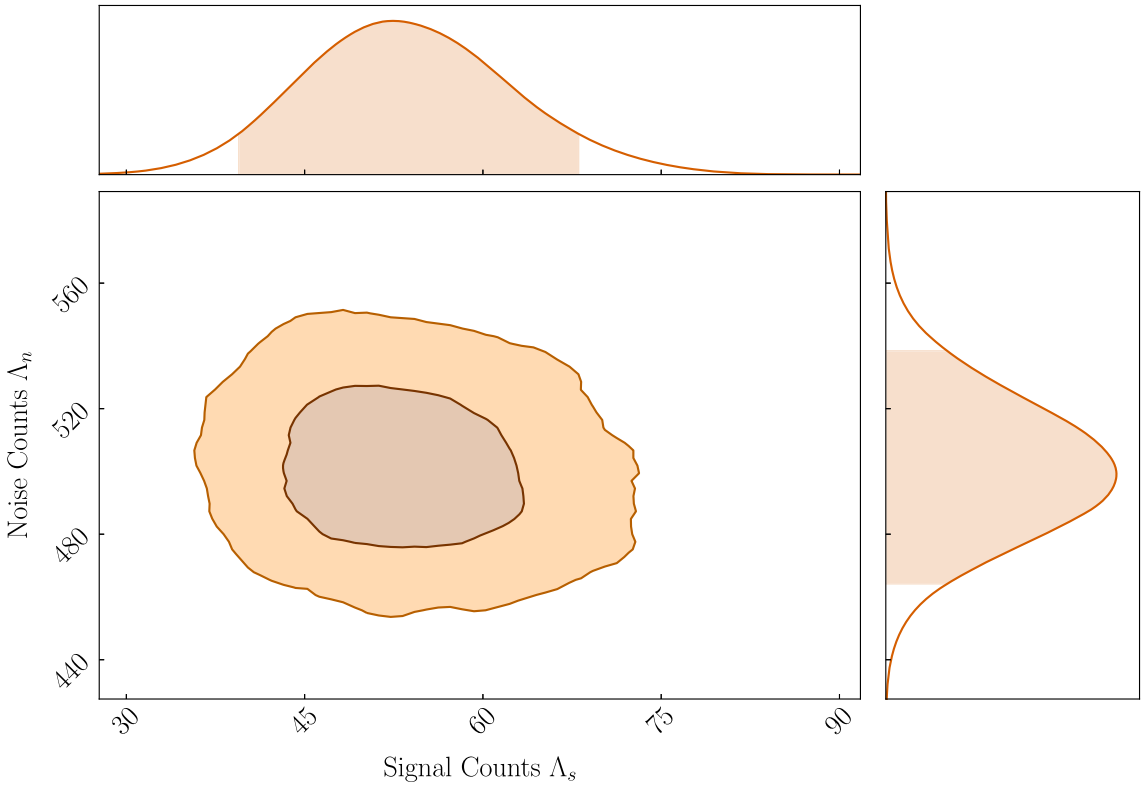


FIG. 5. The posterior for the signal and noise counts in the joint analysis. The shaded region in the one-dimensional posterior corresponds to 90% uncertainty levels, while the contours in the two-dimensional posteriors are the 50% and 90% levels. We recover a median value $\Lambda_s = 53^{+10}_{-13}$.

TABLE I. Triggers with a unified $p_{\rm astro} \geq 0.5$ from our illustrative analysis. The triggers that have $p_{\rm astro} \geq 0.5$ in at least one pipeline in GWTC-2.1 [19] are shown in the second column. Also listed are the FARs of the triggers from the GstLAL and PyCBC pipelines. The GW name of the triggers also recovered with $p_{\rm astro} \geq 0.5$ by at least one pipeline in GWTC-2.1 is given in the second column. These results illustrate the properties of the unified $p_{\rm astro}$ method, but a larger number of noise triggers, and more accurate population models, would be needed to obtain reliable quantitative results.

		GstLAL		PyCBC		
GPS time	GW name	FAR (yr^{-1})	β	FAR (yr^{-1})	β	Unified $p_{\rm astro}$
1238782700.0	GW190408_181802	2.1×10^{-15}	38.4	2.5×10^{-04}	12.9	>0.99
1239082262.0	GW190412_053044	1.9×10^{-27}	66.1	1.1×10^{-04}	13.7	>0.99
1240327333.0	…	9.1×10^{-01}	4.7	$4.2 \times 10^{+01}$	1.2	0.98
1240944862.0	GW190503_185404	2.3×10^{-06}	17.6	3.8×10^{-02}	7.9	>0.99
1241108686.0	…	$1.7 \times 10^{+01}$	1.9	…	…	0.85
1241719652.0	GW190512_180714	7.7×10^{-12}	30.2	1.1×10^{-04}	13.7	>0.99
1241816086.0	GW190513_205428	1.3×10^{-05}	15.8	…	…	>0.99
1242107479.0	GW190517_055101	4.5×10^{-03}	10.0	9.4×10^{-03}	9.3	>0.99
1242315362.0	GW190519_153544	2.2×10^{-06}	17.6	1.0×10^{-04}	13.8	>0.99
1242442967.0	GW190521_030229	2.0×10^{-01}	6.2	4.4×10^{-01}	5.4	>0.99
1242459857.0	GW190521_074359	5.0×10^{-33}	79.0	1.8×10^{-05}	15.5	>0.99
1242984073.0	GW190527_092055	2.3×10^{-01}	6.1	…	…	>0.99
1243533585.0	GW190602_175927	1.1×10^{-07}	20.6	2.9×10^{-01}	5.9	>0.99
1243926576.0	…	$1.2 \times 10^{+01}$	2.2	$6.4 \times 10^{+02}$	0.1	0.72
1243985856.0	…	$3.2 \times 10^{+02}$	0.3	$2.6 \times 10^{+02}$	0.3	0.59
1245221073.0	…	$1.2 \times 10^{+02}$	0.6	$1.8 \times 10^{+02}$	0.4	0.53
1245874666.0	…	$6.2 \times 10^{+02}$	0.1	$2.2 \times 10^{+02}$	0.4	0.80
1246048404.0	GW190701_203306	5.7×10^{-03}	9.8	6.4×10^{-02}	7.4	>0.99
1246385767.0	…	$5.3 \times 10^{+00}$	3.0	…	…	>0.99
1246487219.0	GW190706_222641	5.0×10^{-05}	14.5	3.7×10^{-04}	12.5	>0.99
1246779793.0	…	$1.0 \times 10^{+01}$	2.4	…	…	0.99
1246849694.0	…	$2.7 \times 10^{+00}$	3.6	…	…	>0.99
1247616534.0	GW190720_000836	4.4×10^{-08}	21.5	1.4×10^{-04}	13.5	>0.99
1248242631.0	GW190727_060333	2.7×10^{-10}	26.6	5.6×10^{-03}	9.8	>0.99
1248331528.0	GW190728_064510	5.4×10^{-16}	39.8	8.2×10^{-05}	14.0	>0.99
1248617394.0	GW190731_140936	3.3×10^{-01}	5.7	…	…	>0.99
1249479778.0	…	$6.6 \times 10^{+00}$	2.8	…	…	>0.99
1250620378.0	…	$5.2 \times 10^{+00}$	3.0	…	…	>0.99
1251009263.0	GW190828_063405	5.0×10^{-27}	65.2	8.5×10^{-05}	14.0	>0.99
1251010527.0	GW190828_065509	3.5×10^{-05}	14.9	2.8×10^{-04}	12.8	>0.99
1251588283.0	…	$1.6 \times 10^{+01}$	2.0	…	…	0.92
1251926900.0	…	$1.0 \times 10^{+01}$	2.4	…	…	>0.99
1252627040.0	GW190915_235702	7.8×10^{-06}	16.4	6.8×10^{-04}	11.9	>0.99
1252699636.0	GW190916_200658	$1.2 \times 10^{+01}$	2.2	…	…	0.99
1252756008.0	GW190917_114630	6.6×10^{-01}	5.0	…	…	>0.99
1252939489.0	…	$9.0 \times 10^{+00}$	2.5	…	…	>0.99
1252987339.0	…	$5.1 \times 10^{+01}$	1.1	$8.0 \times 10^{+01}$	0.8	0.65
1253326744.0	GW190924_021846	5.0×10^{-10}	26.0	8.2×10^{-05}	14.0	>0.99
1253509434.0	GW190926_050336	$1.1 \times 10^{+00}$	4.5	…	…	>0.99
1253755327.0	GW190929_012149	1.5×10^{-01}	6.5	$1.2 \times 10^{+02}$	0.6	>0.99
1253885759.0	GW190930_133541	4.3×10^{-01}	5.5	1.8×10^{-02}	8.6	>0.99

Using the signal and noise models described above, we can now proceed to calculate a unified $p_{\rm astro}$ for these triggers.

First, through Eq. (10), we estimate $\Lambda_{\rm s}$ and $\Lambda_{\rm n}$ for the joint analysis. We do this using the Markov-chain Monte Carlo sampler EMCEE [60]. We use uniform priors on the count parameters, $\Lambda_{\rm s} \in [0, 1000]$ and $\Lambda_{\rm n} \in [0, 1000]$. The triggers are then scored against the appropriate distribution to calculate $p(x_i|\mathcal{S})$ and $p(x_i|\varnothing)$.

Figure 5 shows the posterior distribution of $\Lambda_{\rm s}$ and $\Lambda_{\rm n}$, with a median $\Lambda_{\rm s}$ of 53. Table I lists the 41 triggers that

TABLE II. Triggers with a unified $p_{\text{astro}}<0.5$ in our illustrative analysis, but with a $p_{\text{astro}} \geq 0.5$ in either GstLAL or PyCBC in GWTC-2.1 [19].

		GstLAL		PyCBC		
GPS time	GW name	FAR (yr^{-1})	β	FAR (yr^{-1})	β	Unified p_{astro}
1239917954.0	GW190421_213856	2.8×10^{-03}	10.5	$5.9 \times 10^{+00}$	2.9	<0.01
1246527224.0	GW190707_093326	2.7×10^{-15}	38.2	9.7×10^{-06}	16.1	<0.01
1248834439.0	GW190803_022701	7.3×10^{-02}	7.2	$8.1 \times 10^{+01}$	0.8	<0.01

have unified $p_{\text{astro}} \geq 0.5$. Most triggers are reported by both GstLAL and PyCBC, but we do have a few triggers that are detected by GstLAL alone, which corresponds to the high β tail of the on-source triggers in the top panel of Fig. 4. This is qualitatively consistent with the 44 triggers found with a $p_{\text{astro}} \geq 0.5$ by at least one pipeline in GWTC-2.1 analysis [2].

In Table I, we recover 26 triggers that are also reported with $p_{\text{astro}} > 0.5$ in at least one search pipeline in the GWTC-2.1 analysis [2] (the corresponding GW identification of these triggers has been given for easy identification) while the remaining 15 triggers were below this threshold. Inspecting the β values of these triggers show that they are usually low, and comparing with the noise distributions, it is plausible that they could be consistent with noise if some of the modeling assumptions are relaxed. The simplified modeling of the noise distribution is expected to lead to the promotion of some noise triggers, while suppressing some real signals.

In the case of joint triggers there is a heavy weight in favor of signal triggers as discussed in Sec. V. The exact quantitative results might be susceptible to small-number statistics, and it is plausible that some of these would be down weighted if we had a more complete noise distribution from more time slides. However, the results show the expected behavior that having multiple pipelines find a candidate when their noise distributions are largely uncorrelated increases our certainty that a candidate is real.

In Table II, we show the list of triggers with a unified $p_{\text{astro}} < 0.5$, but with a $p_{\text{astro}} \geq 0.5$ in either GstLAL or PyCBC in GWTC-2.1 [19]. While the correlation between pipelines increases p_{astro} for some candidates, the unified analysis also down weights certain triggers, in particular those with highly asymmetric β values. This is because the joint simulated-signal trigger distribution has a strong correlation between the β (and thereby the FAR) values of two pipelines; this is expected since both are matched filter pipelines using CBC templates. For example, GW190803_022701 has GstLAL and PyCBC FARs of 7.3×10^{-2} yr^{-1} and 81 yr^{-1}, respectively. Therefore, while the GstLAL significance of the trigger is high, the asymmetric FARs in the two pipelines gives it a lower weight in the unified analysis. In our case, the low p_{astro} can potentially be explained by the choice of a Gaussian for the noise distribution with heavy tails at large β. The uncertainties involved in the noise and the signal distributions in this illustrative analysis, and the fact that we do not include other search pipelines, mean that we should not rule out the triggers in Table II. Nevertheless, the results demonstrate that there is information that is uniquely captured by a joint analysis of multiple pipelines. In practice, we would expect that any high-significance candidate that is a clear outlier in the signal distribution would warrant further investigation to understand why the pipelines differ in their response.

Finally, many of the triggers in Table I have a p_{astro} close to 1. This is likely an overestimate, due the sparseness of the noise distribution which can lead to an artificially low noise likelihood in Eq. (15). We equally expect that some of the low-probability candidates, not shown in the table, have underestimated p_{astro} for the same reason. The specific values we get are also dependent on the strong modeling assumptions made for the noise distribution. A more realistic analysis would probably require adequate nonparametric modeling of the noise distribution such that any structure in the parameter space can be identified. While it is worth bearing these limitations in mind, both Table I and Fig. 5 qualitatively demonstrate consistency of the results with GWTC-2.1, indicating that even with simpler modeling choices this method can yield sensible overall results.

VI. DISCUSSION AND CONCLUSION

We have developed a statistical formalism for combining information from multiple search pipelines to calculate a unified p_{astro}. We first demonstrated this formalism using a simple toy model, showing that it can consistently combine information and even account for biases in search pipelines. We then applied the framework to O3a data, using triggers from GstLAL and PyCBC, demonstrating how correlations between pipelines may update our understanding of candidates, but highlighting the importance of using accurate models for the signal and noise populations.

Currently, multiple search pipelines are used to identify interesting candidates, each with their own strengths. A unified p_{astro} can potentially reduce confusion in interpretation of triggers found by different pipelines, and can help in using marginal triggers for subsequent GW analyses. We have shown that certain types of correlations between pipelines can inform the significance that the unified analysis assigns to a trigger. Similarly, combining the

search pipelines' results mitigates the need to calculate an effective trials factor to correct the individual-pipeline FARs to account for repeated analysis of the same data. Calculating this trials factor is generally nontrivial. Our formalism can naturally and consistently account for these considerations since it calculates how triggers in one pipeline are correlated with those in another pipeline.

While we have demonstrated that this method can estimate signal and noise counts, and a unified $p_{\rm astro}$ in a sensible and consistent manner, the paucity of noise triggers is a clear computational bottleneck. The scarcity of these triggers can be understood by considering the FAR thresholds used. The 2 day^{-1} threshold used means that we will be limited to few hundred noise triggers in each pipeline per a six month run. Therefore, depending on the structure in the joint noise correlations, we would probably need $\mathcal{O}(10\text{–}100)$ time slides to obtain an accurate representation of the noise distribution. The number of noise triggers that are common between pipelines will be only a fraction of the total number of noise triggers, and if the noise backgrounds are distinct, then there will only be a small fraction in common. This may indicate a need for multiple coordinated time shifts between pipelines to build up the noise distribution. One can also consider other methods of building noise distributions such as modeling the joint noise distribution and drawing from it, similar in philosophy to the GstLAL pipeline [47].

However, while the low number of common noise triggers does pose a difficulty for reconstructing the shape of the noise distribution, it also highlights the importance of considering the number of candidates in common between pipelines: if common noise triggers are rare, but common signal triggers are frequent, then a candidate being found by multiple pipelines should increase its $p_{\rm astro}$. We already see this in the analysis done here where a joint trigger is about 100 times more likely under the signal hypothesis.

If suitable data products (injection sets and noise triggers) are available, this formalism can be readily applied to include other search pipelines, notably those used by the LVK such as MBTA [36,37] and cWB [39–41], as well as pipelines developed for external analysis of public data [61]. This would enable the construction of a single GW catalog accounting for all analyses. A strength of this formalism when constructing joint catalogs is that it would differentially weigh search pipelines, depending on their precision and accuracy by taking into account the shape of the joint simulated-signal distribution. Hence, as long as the simulated-signal distribution sufficiently reflects the true distribution of GW sources, it should be possible to correct for pipelines missing signals over some region of parameter space, or producing spurious triggers in another.

The addition of cWB (and other minimally modeled pipelines [62,63]) to catalog production could be extremely useful, as it can be sensitive in regions of the CBC parameter space where modeled searches do not perform well, such as eccentric binaries [64], in addition to non-CBC transient sources. The sensitivity to non-CBC sources is a complication for current analyses, as the $p_{\rm astro}$ calculations are done assuming only CBC sources (discussed in Appendix F of the GWTC-3 paper [18]). Non-CBC source have a lower true alarm rate than CBC sources, and hence a lower $p_{\rm astro}$ at a given FAR; consequently, misidentifying a trigger as CBC will lead to an overestimate of its $p_{\rm astro}$. To mitigate this, LVK analyses have imposed an additional criterion for cWB candidates, that they must have a counterpart trigger from a template-based CBC pipeline [2,14]. This is effectively an approximation to the unified $p_{\rm astro}$ framework, acknowledging that for real CBC signals there would probably be correlation between pipeline results. This could be put on a more rigorous basis by explicitly using out unified $p_{\rm astro}$ framework, considering the response of the pipelines to simulated CBC and non-CBC signals.

Addition of more pipelines can make the problem of noise triggers more acute, as the number of triggers that are common between three or more pipelines might be small and extrapolation difficult. However, it is probable that noise triggers in a multiple-pipeline space would be so rare that we can conservatively set the noise likelihood to a small limiting value for such a case.

In this paper, we have only considered the simplest case of calculating a unified $p_{\rm astro}$. In a future work, we will extend it to implement binning in the mass space to estimate the astrophysical probability that a system is a BBH, a BNS or a neutron star-black hole binary (NSBH) [25]. Such an extension is in principle straightforward, where in addition to a statistic like the FAR we also fit distribution of recovered template chirp masses for simulated BBH, BNS and NSBH signal. On-source triggers can then be scored against such two-dimensional distributions to give the corresponding likelihood. Searches are often optimized in different ways giving them differential sensitivity in specific parts of the mass space (e.g., PyCBC has an analysis tuned to BBHs [18,49]). Calculating unified probabilities in conjugation with mass binning will allow us to fold in this differential sensitivity in a consistent, unbiased way in a way akin to toy model example with pipeline 3 and pipeline 4 in Sec. IV.

A key uncertainty in calculation of $p_{\rm astro}$ is the form of the underlying source population. Errors in the assumed population translates to a misestimation of the true alarm rate, and hence $p_{\rm astro}$. A way to mitigate this would be to infer the population simultaneously to calculating $p_{\rm astro}$ [44,45,65]. This additionally enables lower significance candidates to be used in population inference. Using a unified $p_{\rm astro}$ makes it easier to assess the contamination fraction in a catalog based upon several pipelines, and hence would make it more convenient to perform such joint inference in the future.

While we have only considered application of our method to final search analyses performed offline, an

extension to low-latency detections is also theoretically possible. By combining multiple pipelines we depend less on one pipeline and hence should be less susceptible to incorrect alerts and retractions. In combination with mass binning, this could be extremely useful for electromagnetic follow-up of GW candidates [66]. Since many such follow-ups are often target-of-opportunity observations, improving the reliability of trigger information can be valuable in evaluating the proper usage of scarce telescope time. The relative scarcity of noise triggers could be a bigger computational issue in low latency, as the joint noise distributions will have to be continuously reevaluated at a reasonable cadence in order to account for the changing detector state [22,23]. Therefore, further work would be needed to identify methods that could potentially be used to reconstruct the noise distribution in low latency.

Finally, Bayesian frameworks have been developed to assess the probability of multimessenger detections [67–70]. These calculate the probabilities associated with candidates being background noise or astrophysical signals from different sources or the same source. Our unified p_{astro} framework naturally feeds into these calculations. Furthermore, our framework would enable extension of these calculations to consider how a nondetection in one messenger impacts the probability that a candidate in a counterpart messenger is real. Given the rich science that may results from a multimessenger discovery [6,71,72], this may be a valuable avenue of future investigation.

The data used in this paper for the analysis of GWTC-2.1 triggers is available as a Zenodo repository [73].

ACKNOWLEDGMENTS

We thank Will Farr, Vicky Kalogera and Surabhi Sachdev for useful discussions. We also thank Anarya Ray, Tom Dent and the anonymous referees for helpful comments on the manuscript. S..B was supported by National Science Foundation (NSF) Grant No. PHY-2207945. C. P. L. B. acknowledges support from Science and Technology Facilities Council (STFC) Grant No. ST/V005634/1. Z. D. acknowledges support from the CIERA Board of Visitors Research Professorship. L. T. is supported by the NSF through Grants No. OAC-2103662 and No. PHY-2011865. G. S. C. D. acknowledges the STFC for funding through Grants No. ST/T000333/1 and No. ST/V005715/1. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by NSF Grants No. PHY-0757058 and No. PHY-0823459. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. This research has made use of data obtained from the Gravitational Wave Open Science Center ([74]) [75], a service of LIGO Laboratory, the LIGO Scientific Collaboration, the Virgo Collaboration, and KAGRA. LIGO Laboratory and Advanced LIGO [3] are funded by the United States NSF as well as the STFC of the United Kingdom, the Max-Planck-Society (MPS), and the State of Niedersachsen/Germany for support of the construction of Advanced LIGO and construction and operation of the GEO 600 detector [76]. Additional support for Advanced LIGO was provided by the Australian Research Council. Virgo [4] is funded, through the European Gravitational Observatory (EGO), by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale di Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by institutions from Belgium, Germany, Greece, Hungary, Ireland, Japan, Monaco, Poland, Portugal, Spain. KAGRA [77] is supported by Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan Society for the Promotion of Science (JSPS) in Japan; National Research Foundation (NRF) and Ministry of Science and ICT (MSIT) in Korea; Academia Sinica (AS) and National Science and Technology Council (NSTC) in Taiwan. Corner plots were made with the ChainConsumer [78] package.

[1] B. P. Abbott *et al.* (LIGO Scientific, Virgo Collaborations), Observation of gravitational waves from a binary black hole merger, Phys. Rev. Lett. **116**, 061102 (2016).
[2] R. Abbott *et al.* (LIGO Scientific, Virgo, KAGRA Collaborations), GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run, arXiv:2111.03606 [Phys. Rev. X (to be published)].
[3] J. Aasi *et al.* (LIGO Scientific Collaboration), Advanced LIGO, Classical Quantum Gravity **32**, 074001 (2015).
[4] F. Acernese *et al.* (Virgo Collaboration), Advanced Virgo: A second-generation interferometric gravitational wave detector, Classical Quantum Gravity **32**, 024001 (2015).
[5] Matthew Pitkin, Stuart Reid, Sheila Rowan, and Jim Hough, Gravitational wave detection by interferometry (ground and space), Living Rev. Relativity **14**, 5 (2011).
[6] B. P. Abbott *et al.* (KAGRA, LIGO Scientific, Virgo Collaborations), Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA, Living Rev. Relativity **23**, 3 (2020).
[7] Benjamin P Abbott *et al.* (LIGO Scientific, Virgo Collaborations), A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals, Classical Quantum Gravity **37**, 055002 (2020).

[8] B. P. Abbott *et al.* (LIGO Scientific, Virgo Collaborations), GW150914: First results from the search for binary black hole coalescence with Advanced LIGO, Phys. Rev. D **93**, 122003 (2016).

[9] B. P. Abbott *et al.* (LIGO Scientific, Virgo Collaborations), Observing gravitational-wave transient GW150914 with minimal assumptions, Phys. Rev. D **93**, 122004 (2016); **94**, 069903(A) (2016).

[10] Will M. Farr, Jonathan R. Gair, Ilya Mandel, and Curt Cutler, Counting and confusion: Bayesian rate estimation with multiple populations, Phys. Rev. D **91**, 023005 (2015).

[11] F. Guglielmetti, R. Fischer, and V. Dose, Background-source separation in astronomical images with Bayesian probability theory (I): The method, Mon. Not. R. Astron. Soc. **396**, 165 (2009).

[12] B. P. Abbott *et al.* (LIGO Scientific, Virgo Collaborations), The rate of binary black hole mergers inferred from Advanced LIGO observations surrounding GW150914, Astrophys. J. Lett. **833**, L1 (2016).

[13] B. P. Abbott *et al.* (LIGO Scientific, Virgo Collaborations), Supplement: The rate of binary black hole mergers inferred from Advanced LIGO observations surrounding GW150914, Astrophys. J. Suppl. Ser. **227**, 14 (2016).

[14] B. P. Abbott *et al.* (LIGO Scientific, 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 (2019).

[15] R. Abbott *et al.* (LIGO Scientific, Virgo, KAGRA Collaborations), The population of merging compact binaries inferred using gravitational waves through GWTC-3, Phys. Rev. X **13**, 011048 (2023).

[16] Alexander H. Nitz, Sumit Kumar, Yi-Fan Wang, Shilpa Kastha, Shichao Wu, Marlin Schäfer, Rahul Dhurkunde, and Collin D. Capano, 4-OGC: Catalog of gravitational waves from compact-binary mergers, arXiv:2112.06878.

[17] Seth Olsen, Tejaswi Venumadhav, Jonathan Mushkin, Javier Roulet, Barak Zackay, and Matias Zaldarriaga, New binary black hole mergers in the LIGO-Virgo O3a data, Phys. Rev. D **106**, 043009 (2022).

[18] R. Abbott *et al.* (LIGO Scientific, Virgo Collaborations), GWTC-2: Compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run, Phys. Rev. X **11**, 021053 (2021).

[19] R. Abbott *et al.* (LIGO Scientific, Virgo Collaborations), 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 [Phys. Rev. D (to be published)].

[20] B. P. Abbott *et al.* (LIGO Scientific, Virgo Collaborations), GW170817: Observation of gravitational waves from a binary neutron star inspiral, Phys. Rev. Lett. **119**, 161101 (2017).

[21] R. Abbott *et al.* (LIGO Scientific, KAGRA, Virgo Collaborations), Observation of gravitational waves from two neutron star–black hole coalescences, Astrophys. J. Lett. **915**, L5 (2021).

[22] Aaron Buikema *et al.* (aLIGO Collaboration), Sensitivity and performance of the Advanced LIGO detectors in the third observing run, Phys. Rev. D **102**, 062003 (2020).

[23] F. Acernese *et al.* (Virgo Collaboration), The Virgo O3 run and the impact of the environment, Classical Quantum Gravity **39**, 235009 (2022).

[24] B. P. Abbott *et al.* (LIGO Scientific, Virgo Collaborations), GW150914: The Advanced LIGO detectors in the era of first discoveries, Phys. Rev. Lett. **116**, 131103 (2016).

[25] Shasvath J. Kapadia *et al.*, A self-consistent method to estimate the rate of compact binary coalescences with a Poisson mixture model, Classical Quantum Gravity **37**, 045007 (2020).

[26] Cody Messick *et al.*, Analysis framework for the prompt discovery of compact binary mergers in gravitational-wave data, Phys. Rev. D **95**, 042001 (2017).

[27] Surabhi Sachdev *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.

[28] Chad Hanna *et al.*, Fast evaluation of multidetector consistency for real-time gravitational wave searches, Phys. Rev. D **101**, 022003 (2020).

[29] Kipp Cannon *et al.*, GstLAL: A software framework for gravitational wave discovery, SoftwareX **14**, 100680 (2021).

[30] Bruce Allen, Warren G. Anderson, Patrick R. Brady, Duncan A. Brown, and Jolien D. E. Creighton, FINDCHIRP: An algorithm for detection of gravitational waves from inspiraling compact binaries, Phys. Rev. D **85**, 122006 (2012).

[31] Tito Dal Canton *et al.*, Implementing a search for aligned-spin neutron star-black hole systems with advanced ground based gravitational wave detectors, Phys. Rev. D **90**, 082004 (2014).

[32] Samantha A. Usman *et al.*, The PyCBC search for gravitational waves from compact binary coalescence, Classical Quantum Gravity **33**, 215004 (2016).

[33] Alexander H. Nitz, Thomas Dent, Tito Dal Canton, Stephen Fairhurst, and Duncan A. Brown, Detecting binary compact-object mergers with gravitational waves: Understanding and improving the sensitivity of the PyCBC search, Astrophys. J. **849**, 118 (2017).

[34] Gareth S. Davies, Thomas Dent, Márton Tápai, Ian Harry, Connor McIsaac, and Alexander H. Nitz, Extending the PyCBC search for gravitational waves from compact binary mergers to a global network, Phys. Rev. D **102**, 022004 (2020).

[35] Tom Dent, Technical note: Extending the PyCBC pastro calculation to a global network, https://dcc.ligo.org/LIGO-T2100060/public.

[36] T. Adams, D. Buskulic, V. Germain, G. M. Guidi, F. Marion, M. Montani, B. Mours, F. Piergiovanni, and G. Wang, Low-latency analysis pipeline for compact binary coalescences in the advanced gravitational wave detector era, Classical Quantum Gravity **33**, 175012 (2016).

[37] F. Aubin *et al.*, The MBTA pipeline for detecting compact binary coalescences in the third LIGO–Virgo observing run, Classical Quantum Gravity **38**, 095004 (2021).

[38] Nicolas Andres *et al.*, Assessing the compact-binary merger candidates reported by the MBTA pipeline in the LIGO–Virgo O3 run: Probability of astrophysical origin, classification, and associated uncertainties, Classical Quantum Gravity **39**, 055002 (2022).

[39] Sergey Klimenko and Guenakh Mitselmakher, A wavelet method for detection of gravitational wave bursts, Classical Quantum Gravity **21**, S1819 (2004).
[40] S. Klimenko, G. Vedovato, M. Drago, G. Mazzolo, G. Mitselmakher, C. Pankow, G. Prodi, V. Re, F. Salemi, and I. Yakushin, Localization of gravitational wave sources with networks of advanced detectors, Phys. Rev. D **83**, 102001 (2011).
[41] S. Klimenko *et al.*, Method for detection and reconstruction of gravitational wave transients with networks of advanced detectors, Phys. Rev. D **93**, 042004 (2016).
[42] Patrick J. Sutton, Upper limits from counting experiments with multiple pipelines, Classical Quantum Gravity **26**, 245007 (2009).
[43] Rahul Biswas *et al.*, Detecting transient gravitational waves in non-Gaussian noise with partially redundant analysis methods, Phys. Rev. D **85**, 122009 (2012).
[44] Shanika Galaudage, Colm Talbot, and Eric Thrane, Gravitational-wave inference in the catalog era: Evolving priors and marginal events, Phys. Rev. D **102**, 083026 (2020).
[45] Javier Roulet, Tejaswi Venumadhav, Barak Zackay, Liang Dai, and Matias Zaldarriaga, Binary black hole mergers from LIGO/Virgo O1 and O2: Population inference combining confident and marginal events, Phys. Rev. D **102**, 123022 (2020).
[46] S. Babak *et al.*, Searching for gravitational waves from binary coalescence, Phys. Rev. D **87**, 024033 (2013).
[47] Kipp Cannon, Chad Hanna, and Drew Keppel, Method to estimate the significance of coincident gravitational-wave observations from compact binary coalescence, Phys. Rev. D **88**, 024025 (2013).
[48] Tejaswi Venumadhav, Barak Zackay, Javier Roulet, Liang Dai, and Matias Zaldarriaga, New binary black hole mergers in the second observing run of Advanced LIGO and Advanced Virgo, Phys. Rev. D **101**, 083030 (2020).
[49] Alexander H. Nitz, Thomas Dent, Gareth S. Davies, Sumit Kumar, Collin D. Capano, Ian Harry, Simone Mozzon, Laura Nuttall, Andrew Lundgren, and Márton Tápai, 2-OGC: Open gravitational-wave catalog of binary mergers from analysis of public Advanced LIGO and Virgo data, Astrophys. J. **891**, 123 (2020).
[50] Alexander H. Nitz, Collin D. Capano, Sumit Kumar, Yi-Fan Wang, Shilpa Kastha, Marlin Schäfer, Rahul Dhurkunde, and Miriam Cabero, 3-OGC: Catalog of gravitational waves from compact-binary mergers, Astrophys. J. **922**, 76 (2021).
[51] Benjamin P. Abbott *et al.* (LIGO Scientific, Virgo Collaborations), Search for intermediate mass black hole binaries in the first observing run of Advanced LIGO, Phys. Rev. D **96**, 022001 (2017).
[52] Rich Abbott *et al.* (LIGO Scientific, Virgo, KAGRA Collaborations), Search for intermediate-mass black hole binaries in the third observing run of Advanced LIGO and Advanced Virgo, Astron. Astrophys. **659**, A84 (2022).
[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, SCIKIT-LEARN: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011), https://jmlr.org/papers/v12/pedregosa11a.html.
[54] Gareth S. Cabourn Davies and Ian W. Harry, Establishing significance of gravitational-wave signals from a single observatory in the PyCBC offline search, Classical Quantum Gravity **39**, 215012 (2022).
[55] LIGO Scientific, Virgo, and KAGRA Collaborations, GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run —O3 search sensitivity estimates (2021), https://zenodo.org/records/5546676.
[56] P. A. R. Ade *et al.* (Planck Collaboration), Planck 2015 results. XIII. Cosmological parameters, Astron. Astrophys. **594**, A13 (2016).
[57] Thomas A. Callister and Will M. Farr, A parameter-free tour of the binary black hole population, arXiv:2302.07289.
[58] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability (Taylor & Francis, London, 1986), https://www.taylorfrancis.com/books/mono/10.1201/9781315140919/density-estimation-statistics-data-analysis-bernard-silverman.
[59] Jam Sadiq, Thomas Dent, and Mark Gieles, Binary vision: The merging black hole binary mass distribution via iterative density estimation, arXiv:2307.12092.
[60] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman, EMCEE: The MCMC hammer, Publ. Astron. Soc. Pac. **125**, 306 (2013).
[61] Tejaswi Venumadhav, Barak Zackay, Javier Roulet, Liang Dai, and Matias Zaldarriaga, 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 (2019).
[62] Ryan Lynch, Salvatore Vitale, Reed Essick, Erik Katsavounidis, and Florent Robinet, Information-theoretic approach to the gravitational-wave burst detection problem, Phys. Rev. D **95**, 104046 (2017).
[63] Jonah B. Kanner, Tyson B. Littenberg, Neil Cornish, Meg Millhouse, Enia Xhakaj, Francesco Salemi, Marco Drago, Gabriele Vedovato, and Sergey Klimenko, Leveraging waveform complexity for confident detection of gravitational waves, Phys. Rev. D **93**, 022002 (2016).
[64] B. P. Abbott *et al.* (LIGO Scientific, Virgo Collaborations), Search for eccentric binary black hole mergers with Advanced LIGO and Advanced Virgo during their first and second observing runs, Astrophys. J. **883**, 149 (2019).
[65] Sebastian M. Gaebel, John Veitch, Thomas Dent, and Will M. Farr, Digging the population of compact binary mergers out of the noise, Mon. Not. R. Astron. Soc. **484**, 4008 (2019).
[66] Ryan Lynch, Michael Coughlin, Salvatore Vitale, Christopher W. Stubbs, and Erik Katsavounidis, Observational implications of lowering the LIGO-Virgo alert threshold, Astrophys. J. Lett. **861**, L24 (2018).
[67] G. Ashton, E. Burns, T. Dal Canton, T. Dent, H. B Eggenstein, A. B. Nielsen, R. Prix, M. Was, and S. J. Zhu, Coincident detection significance in multimessenger astronomy, Astrophys. J. **860**, 6 (2018).

[68] Imre Bartos, Doga Veske, Azadeh Keivani, Zsuzsa Marka, Stefan Countryman, Erik Blaufuss, Chad Finley, and Szabolcs Marka, Bayesian multi-messenger search method for common sources of gravitational waves and high-energy neutrinos, Phys. Rev. D **100**, 083017 (2019).
[69] Doğa Veske, Zsuzsa Márka, Imre Bartos, and Szabolcs Márka, How to search for multiple messengers—a general framework beyond two messengers, Astrophys. J. **908**, 216 (2021).
[70] Brandon Piotrzkowski, Amanda Baylor, and Ignacio Magaña Hernandez, A joint ranking statistic for multi-messenger astronomical searches with gravitational waves, Classical Quantum Gravity **39**, 085010 (2022).
[71] B. P. Abbott *et al.* (LIGO Scientific, Virgo, Fermi GBM, INTEGRAL, IceCube, AstroSat Cadmium Zinc Telluride Imager Team, IPN, Insight-Hxmt, ANTARES, Swift, AGILE Team, 1M2H Team, Dark Energy Camera GW-EM, DES, DLT40, GRAWITA, Fermi-LAT, ATCA, ASKAP, Las Cumbres Observatory Group, OzGrav, DWF (Deeper Wider Faster Program), AST3, CAASTRO, VINROUGE, MASTER, J-GEM, GROWTH, JAGWAR, CaltechNRAO, TTU-NRAO, NuSTAR, Pan-STARRS, MAXI Team, TZAC Consortium, KU, Nordic Optical Telescope, ePESSTO, GROND, Texas Tech University, SALT Group, TOROS, BOOTES, MWA, CALET, IKI-GW Follow-up, H.E.S.S., LOFAR, LWA, HAWC, Pierre Auger, ALMA, Euro VLBI Team, Pi of Sky, Chandra Team at McGill University, DFN, ATLAS Telescopes, High Time Resolution Universe Survey, RIMAS, RATIR, SKA South Africa/MeerKAT Collaborations), Multi-messenger observations of a binary neutron star merger, Astrophys. J. Lett. **848**, L12 (2017).
[72] Raffaella Margutti and Ryan Chornock, First multimessenger observations of a neutron star merger, Annu. Rev. Astron. Astrophys. **59**, 155 (2021).
[73] https://zenodo.org/records/8432160.
[74] https://gwosc.org/.
[75] R. Abbott *et al.* (LIGO Scientific, Virgo, KAGRA Collaborations), Open data from the third observing run of LIGO, Virgo, KAGRA and GEO, Astrophys. J. Suppl. Ser. **267**, 29 (2023).
[76] K. L. Dooley *et al.*, GEO 600 and the GEO-HF upgrade program: Successes and challenges, Classical Quantum Gravity **33**, 075009 (2016).
[77] T. Akutsu *et al.* (KAGRA Collaboration), Overview of KAGRA: Calibration, detector characterization, physical environmental monitors, and the geophysics interferometer, Prog. Theor. Exp. Phys. **2021**, 05A102 (2021).
[78] S. R. Hinton, ChainConsumer, J. Open Source Software **1**, 45 (2016).

Correction: The previously published running head contained an error in the third author name and has been set right.