Calibrating gravitational-wave search algorithms with conformal prediction

Gregory Ashton,[1,*] Nicolo Colombo,[2] Ian Harry,[3] and Surabhi Sachdev[4]
[1]*Department of Physics, Royal Holloway, Egham TW20 0EX, United Kingdom*
[2]*Department of Computer Science, Royal Holloway, Egham TW20 0EX, United Kingdom*
[3]*University of Portsmouth, Portsmouth PO1 3FX, United Kingdom*
[4]*School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332, USA*

(Received 19 March 2024; accepted 17 May 2024; published 17 June 2024)

In astronomy, we frequently face the decision problem: does this data contain a signal? Typically, a statistical approach is used, which requires a threshold. The choice of threshold presents a common challenge in settings where signals and noise must be delineated, but their distributions overlap. Gravitational-wave astronomy, which has gone from the first discovery to catalogs of hundreds of events in less than a decade, presents a fascinating case study. For signals from colliding compact objects, the field has evolved from a frequentist to a Bayesian methodology. However, the issue of choosing a threshold and validating noise contamination in a catalog persists. Confusion and debate often arise due to the misapplication of statistical concepts, the complicated nature of the detection statistics, and the inclusion of astrophysical background models. We introduce conformal prediction (*CP*), a framework developed in machine learning to provide distribution-free uncertainty quantification to point predictors. We show that *CP* can be viewed as an extension of the traditional statistical frameworks whereby thresholds are calibrated such that the uncertainty intervals are statistically rigorous and the error rate can be validated. Moreover, we discuss how *CP* offers a framework to optimally build a metapipeline combining the outputs from multiple independent searches. We introduce *CP* with a toy cosmic-ray detector, which captures the salient features of most astrophysical search problems and allows us to demonstrate the features of *CP* in a simple context. We then apply the approach to a recent gravitational-wave mock data challenge using multiple search algorithms for compact binary coalescence signals in interferometric gravitational-wave data. Finally, we conclude with a discussion on the future potential of the method for gravitational-wave astronomy.

DOI: 10.1103/PhysRevD.109.123027

I. INTRODUCTION

The burgeoning field of gravitational-wave astronomy is in a state of rapid evolution. Second-generation detectors [1–3] have progressed from the first observation of a binary black hole merger [4] to the compilation of extensive transient event catalogs [5–7] including also binary neutron star and black hole neutron star mergers. With this progress, the methodologies for evaluating the statistical significance of compact binary coalescence (CBC) signals have undergone notable transformations. While the significance of the initial detection [4] was assessed through the frequentist false alarm rate (FAR), contemporary catalogs [5–7] now use probabilistic Bayesian methods.

*gregory.ashton@rhul.ac.uk

However, astrophysicists aiming to learn from gravitational-wave data are confronted with a challenge: the difficulty in identifying signals when their distribution and the noise distributions overlap. This issue is by no means unique in astronomy (see, e.g., Feigelson and Babu [8]). However, gravitational-wave astronomy is an especially intriguing case study because the signal-to-noise ratio (SNR) of sources is low, but the potential scientific reward is high. Moreover, much of the insights derive from studying the population of identified sources [9]. The events producing signals within current sensitivities are isotropically distributed, so the number of detections scales with the cube of the horizon distance (a measure of the detector sensitivity). Therefore, there are always more events just beyond the horizon than within: increasing the horizon distance by just 25% will double the number of events. The conundrum facing anyone wishing to utilize the hundreds of sources now reported is how to select a threshold to cut between the signals and the noise. On the one hand, we can choose a conservative threshold, ensuring a high catalog purity (the fraction of true signals).

However, the conservative threshold also entails a loss of accuracy; after all, we must discard many low-significance astrophysical signals buried in the noise. On the other hand, choosing a liberal threshold would include a larger number of astrophysical signals but at the cost of bias induced by nonastrophysical catalog contamination.

Along with the threshold problem, difficulties arise from concurrently applying multiple search algorithms (hereafter referred to as "pipelines"). The Gravitational-Wave Transient Catalog (GWTC) produced by the LIGO Scientific, Virgo, and KAGRA (LVK) Collaborations (e.g., Abbott *et al.* [5]) include results from several independent pipelines (specifically, GstLAL [10–15], MBTA [16,17], PyCBC [18–22], SPIIR [23,24], and COHERENT WAVEBURST [25]). For a given candidate event, the significance between pipelines can vary substantially, reflecting inherent uncertainty in the significance estimate and varying pipeline performance. However, for those not intimately knowledgeable about the ever-evolving internal workings of the pipelines, it is hard to know when a particular pipeline is more reliable or more sensitive than another.

There are efforts underway to address these issues. For example, population-level analyses can utilize hierarchical models to assess mixed catalogs of signals and noise, avoiding the contamination problem altogether [26–29] and recent efforts are also underway to produce a unified significance estimate [30]. Nevertheless, the problem of choosing thresholds will continue to be of interest as mixed methods are in their infancy, and some of the most interesting events will inevitably come from close to the detection horizon: the question of "do these data contain an astrophysical signal?" will inevitably persist.

This work will introduce a new and transformative framework to solve this problem using conformal prediction (*CP*) [31]. *CP* is an approach to uncertainty quantification developed within the context of machine learning (ML). *CP* takes an existing point-prediction algorithm and a calibration dataset (consisting of correctly labeled data) and generalizes the underlying algorithm's point prediction to a "prediction set" with a guaranteed "validity" (where valid means that the true label is guaranteed to belong to the set with a predefined confidence). Its appeal arises from its universal applicability, guarantees, and single assumption: exchangeability of the data. Moreover, the prediction guarantees are distribution-free: there is no asymptotic assumption or underlying model. It can be used for classification and regression or, correspondingly, search/detection and inference/measurement in the language of astronomy [32]. This work will explore the classification (or search/detection) problem. We will demonstrate how *CP* can be applied to calibrate pipelines without requiring knowledge of its internal behavior. Moreover, we will discuss how *CP* offers an alternative approach to developing a metapipeline: taking the inputs from multiple search algorithms and providing a single statement which optimally combines their outputs and is well calibrated.

As we will show, *CP* is simple to implement, easily tested, has minimal assumptions, and no required astrophysical model. For these reasons, we anticipate that *CP* will be of general interest to the field. While we will discuss *CP* exclusively in the context of searching for CBC signals, we anticipate it will find utility for searches for other sources of gravitational-wave radiation and beyond.

The remainder of this article is structured as follows. In Sec. II, we introduce the existing traditional approaches for significance estimation within gravitational-wave astronomy and further motivate this work by considering their real-world performance. We provide a lay guide to *CP* in Sec. III. We apply it in Sec. IV to a toy cosmic-ray detector problem to demonstrate the basic algorithm and extensions in the noise-dominated regime. Moreover, we also use our toy problem to explain some of the subtleties of *CP*. In Sec. VI, we then go on to apply *CP* to the recent mock data challenge of LIGO-Virgo data [33]. Finally, we end with a discussion on the advantages, difficulties, and future prospects of *CP* for gravitational-wave astronomy in Sec. VII.

II. METHODOLOGY: QUANTIFYING SIGNIFICANCE WITH TRADITIONAL APPROACHES

To begin our discussion, we first review the data, search algorithms, detection statistics, and two dominant quantities used to assess candidate significance: FAR and p_{astro}. Gravitational-wave strain data comprise quasistationary colored Gaussian background noise, astrophysical signals, and a variety of nonastrophysical transient noise sources termed "glitches" [34,35]. Absent glitches, the optimal detection statistic is the colored Gaussian noise matched-filter SNR. When the signal source properties (e.g., the mass of the system) are unknown (as is typical), a bank of templates is searched, often in combination with techniques to maximize or marginalize over subsets of the full parameter space (see, e.g., Sathyaprakash and Dhurandhar [36]). However, in the presence of glitches, the optimal statistic is unknown. To guide the reader on how the leading searches remain sensitive to astrophysical signals despite frequent glitches, we now describe in broad terms a typical search algorithm or pipeline: the interested reader may wish to review Abbott *et al.* [35] for a deeper discussion.

The central tools used by most pipelines to distinguish between signals and glitches are the coincidence between detectors and signal consistency checks such as the χ^2 detection statistic [18], which discriminates cases where the data are likely to contain a glitch by analyzing the way power is distributed in the broadband signal. Typically, the χ^2 and matched-filter SNR are combined to produce a "combined ranking statistic" which we label ρ. Additional terms may also be included in the combined ranking statistics, such as weights based on whether the region

of parameter space is expected to contain more astrophysical signals and amplitude-phase-time consistency checks between detectors. The combined ranking statistic can be tuned to maximize the separation of signals from noise (as verified by simulations). Since the combined ranking statistic is *ad hoc*, its background distribution (where the background is taken to mean in the absence of any astrophysical signal) is inherently unknown and must be empirically estimated from the data. However, gravitational-wave detectors cannot be shielded from astrophysical signals. Therefore, pipelines use approaches such as "time sliding" between separate independent detectors to destroy correlations between astrophysical signals (see, e.g., [37,38]), resulting in empirical measurement of the background. We denote such a background as the set $\{\rho\} = \{\rho_0, \rho_1, \ldots, \rho_{n-1}\}$ of n values measured on the background.

Once the background has been estimated for a new candidate event with ranking statistic ρ', the pipeline estimates its significance by calculating the FAR. Informally, the FAR is the amount of background data one must observe to see a ranking statistic as large as ρ'. Such a dimensionful approach results in an intuitive understanding of the significance given knowledge of the amount of data searched. For example, for a search of one month of data, an event with a FAR of 1 per millennia is a clear detection, while a FAR of 1 per day is more likely to be noise. More precisely, the FAR is calculated empirically as the inverse of the number of background events with a ranking statistic of ρ' or greater divided by the segment duration used in the search. One sees then that the FAR is the one-sided right-tail empirical p-value divided by the segment duration

$$\mathrm{FAR} = \frac{1}{T}\Pr(\rho > \rho' | H_0) = \frac{1}{T}\frac{|\{\rho_i : \rho_i > \rho'\}|}{|\{\rho\}|}, \quad (1)$$

where H_0 is the null hypothesis, we apply set-builder notation, and define the set size by $|\cdot|$.

The FAR of the first detection was reported in the paper abstract: "less than 1 event per 203 000 years" [4]. However, once a population of signals was established, it became preferential to move to a probabilistic approach instead. Following Farr *et al.* [39], the foreground and background distributions are modeled by a Poisson mixture model with prior choices informed by the pipeline outputs and previously observed signals. From this, each pipeline produces a new significance estimate, p_{astro}: the probability that the signal is astrophysical [40–43]. Moreover, the modeled approach allows further subclassification as $p_{\mathrm{astro}} = p_{\mathrm{BNS}} + p_{\mathrm{NSBH}} + p_{\mathrm{BBH}}$ (and a complementary probability of terrestrial origin; BNS, binary neutron star; NSBH, binary neutron star; BBH, binary black hole). With this new approach, GWTC-1 [44] defined "GW" events as those with a FAR less than 1 per 30 days and a p_{astro} greater than $1/2$. This latter definition has become a *de facto* standard. For example, a p_{astro} greater than $1/2$ is the threshold used to identify events for further follow-up in several recent catalogs [5–7]. Yet, it demonstrates that, even with a probabilistic interpretation of the nature of a candidate, researchers still like to establish a threshold and draw a clear delineation, and it is quite common to see astrophysics research take the provided thresholds at face value.

The final complicating piece of this picture is that multiple pipelines analyze the same data. Our typical pipeline above described the core features, but each employs a unique arsenal of techniques built over many years by many people. The result is that, for any given candidate, we end up with multiple estimates of its significance: a FAR and p_{astro} per pipeline. The pipelines broadly agree for unambiguous signals and noise events where apples-to-apples comparisons can be made. However, it is in the gray middle ground where things become complicated. To demonstrate this, we use data from the recent GWTC-3 catalog [5], which reported on data from the second part of the third LIGO-Virgo observing run. We use the associated data release, which includes triggers where at least one pipeline had a FAR of less than 2 per day: as such, we expect this to include both the astrophysical signals and a great number of nonastrophysical candidates.

In Fig. 1, we scatterplot the p_{astro} of each trigger for pairs of CBC search pipelines used in GWTC-3 (we exclude the COHERENT WAVEBURST pipeline that applies an unmodeled search approach). In the off-diagonal corners, two dense

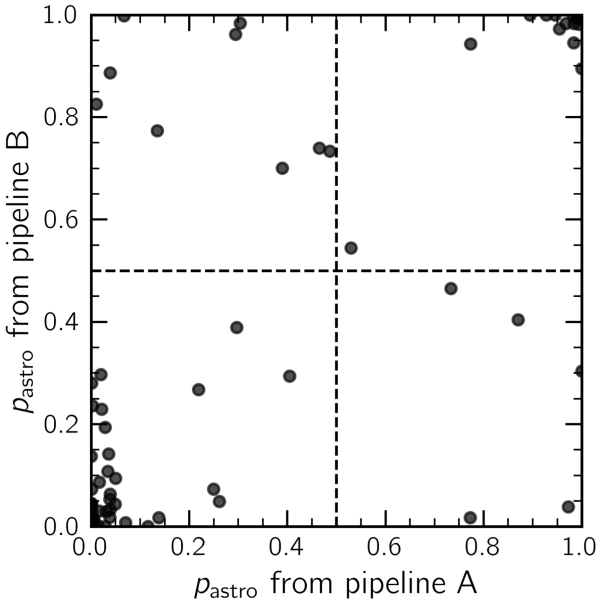

FIG. 1. Comparison of the probability of astrophysical origin estimated by pairs of pipelines for all candidates reported in GWTC-3 (including subthreshold candidates). While clear signal (top right) and clear noise (bottom left) cases usually agree, a significant off-diagonal scatter remains between these points.

regions correspond to the clear signal (top right) and clear noise (bottom left) cases where pipelines agree. However, scattered through the plane are confusion cases where one pipeline finds $p_{\rm astro} > 0.5$, indicating the data contained an astrophysical source, while the other pipeline is more pessimistic ($p_{\rm astro} < 0.5$). If we are lucky enough to know experts from both pipelines, we can understand the cause of the discrepancy. Sometimes, it is well understood different choices lead to different sensitivities in different parts of the parameter space. If the more sensitive pipeline found the event while the other did not, this explains the difference, and we may gain confidence that this is an astrophysical signal. Other times, the differences are more contentious or yet to be understood—this should be expected, as these are complicated multistage pipelines with differing and often implicit assumptions. Nevertheless, it leaves the uninformed with the previously described choice-of-threshold conundrum exacerbated by the need to learn the detailed inner workings of the pipeline to understand the results. One standard solution is to take the maximum $p_{\rm astro}$, implicitly trusting that the only explanation is variations in sensitivity. However, another explanation is random uncertainty in significance or even that one pipeline is malperforming.

One may imagine that the inclusion of different astrophysical foreground prior models in the Bayesian analysis may explain the scatter in Fig. 1 between pipelines; however, Fig. 2 demonstrates that the scatter is also inherent in the underlying and simpler FAR. Finally, in Fig. 3, we plot each pipeline's FAR against $p_{\rm astro}$. Here, we see the approximate sigmoid relationship with significant scatter.

The GWTC-3 results demonstrate the inherent difficulty facing anyone wishing to select a set of events for further analysis. However, these results are only part of the picture. They present only the pipelines used by the LVK Collaborations. There are external groups that produce independent catalogs where the same conclusions hold up: scatter between significance estimates. Moreover, pipelines are not static: they are constantly developed, improved, and reconfigured. It is well known that the same pipeline with a different configuration can produce a different significance estimate (usually for well-understood reasons understood by the pipeline experts). Therefore, even choosing a single pipeline can effectively represent a different pipeline per observing run (or period in which the methodology and configuration are static). Finally, using $p_{\rm astro}$ as a threshold also utilizes information from estimates of the population properties. Since we are constantly learning new information and improving estimates, this can lead to the reranking of past data, resulting in the possibility of reclassifying old candidates.

One naive way of describing the situation is that significance estimates (i.e., the FAR or $p_{\rm astro}$) do not come with an associated uncertainty (from, e.g., intrinsic configuration choices, population choices, or data choices). The oft-used approach to resolve this is to take the scatter from multiple pipelines as a proxy indication of the uncertainty. This has primarily been the community approach: confidence in the first detection from a new source class is validated by the involvement of multiple pipelines. However, this is not satisfactory and discards

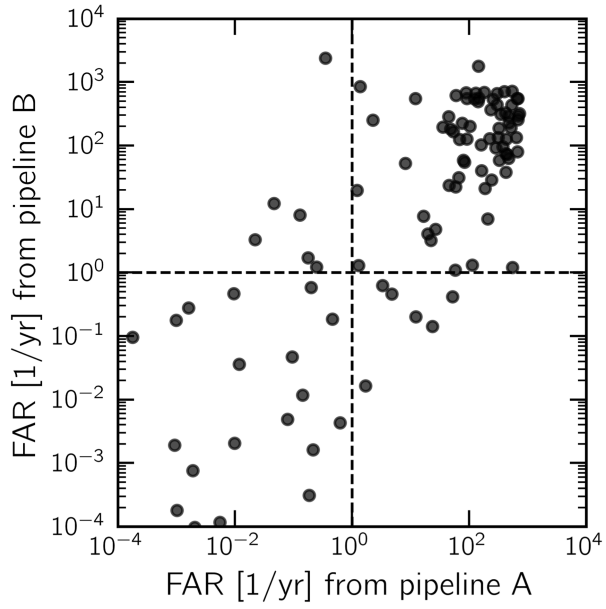

FIG. 2. Comparison of the FAR estimated by pairs of pipelines demonstrating the intrinsic scatter for all candidates reported in GWTC-3 (including subthreshold candidates). While clear signal (bottom left) and clear noise (top right) cases usually agree, a significant off-diagonal scatter remains between these points.

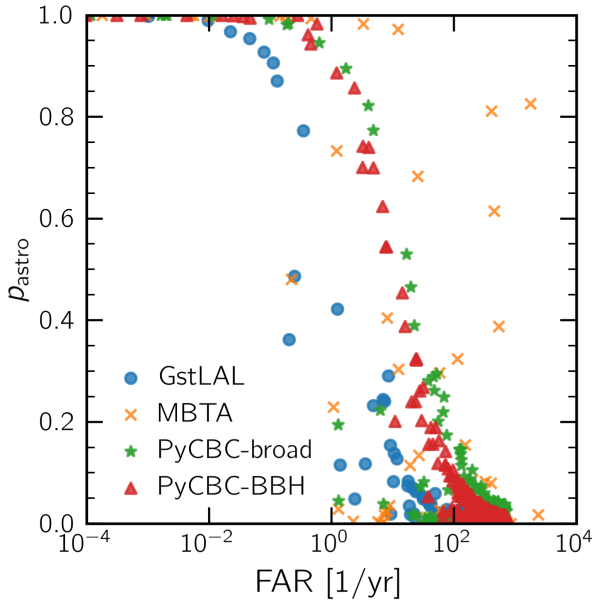

FIG. 3. Comparison by pipeline between $p_{\rm astro}$ and FAR for all candidates reported in GWTC-3 (including subthreshold candidates).

inherent information about pipeline sensitivity. In the remainder of this article, we will introduce a formal alternative based on *CP*. Our fundamental interest is to develop a tool that takes the FAR or $p_{\rm astro}$ as a heuristic and calibrates it, enabling standardization between pipelines and proper uncertainty estimates for whether a candidate is of astrophysical origin.

III. METHODOLOGY: QUANTIFYING SIGNIFICANCE WITH CONFORMAL PREDICTION

We now introduce the *CP* methodology. We intend to give the reader a guide to the application without delving into the foundational theory, which can be found in reviews such as Angelopoulos and Bates [45] and Shafer and Vovk [46].

To begin, it should be understood that *CP* was developed in the machine learning classification algorithm context. Specifically, it can be applied to any classification algorithm, i.e., given some observed data x, an algorithm that produces a single predicted "label $y^{(\ell)}$" drawn from a set of N possible labels $\{y^{(0)}, y^{(1)}, \ldots, y^{(N-1)}\}$. *CP* calibrates the classification algorithm by producing a prediction set Γ^α where $\alpha \in [0, 1]$ is the allowed "error rate," also known as the "significance level." The steps to generate the prediction set are as follows:

(1) Definitions: Define a nonconformity measure $A(x, y^{(\ell)})$, which returns a nonconformity score s for each label in the complete set. The requirements for the nonconformity score are loose; it must simply be a real-valued number. However, for the algorithm to be useful, the score should be large when $y^{(\ell)}$ is not the correct label (i.e., it measures how unusual the labeling would be).

(2) Calibration: Now define the calibration data: n pairs of $(x, y^{(\hat{\ell})})$ where x is the observed data and $y^{(\hat{\ell})}$ is the true label (indicated by the hat on the index). In our context, calibration data will always be drawn from simulations. Now, for each element of the calibration data, calculate the equivalent score for the true label and store this in a set of calibration scores $s_i = A(x_i, y_i^{(\hat{\ell})})$ where the lower subscript i is added to indicate the ith element of the calibration data.

(3) Quantile: The final step before generating the prediction set is to define the allowed error rate $\alpha \in [0, 1]$, then given a set of calibration scores, we calculate

$$\hat{q} = s_{(\lceil (n+1)(1-\alpha) \rceil)}, \tag{2}$$

where $\lceil \cdot \rceil$ is the ceiling function, and we indicate by the use of $s_{(j)}$ the jth value of the ordered set of s_i. As described in Angelopoulos and Bates [45], $\hat{q}$ is essentially the $1-\alpha$ quantile of the calibration scores with a small correction.

(4) Prediction: Finally, given a new observed data point x', we generate the prediction set

$$\Gamma^\alpha = \left\{ y^{(\ell)} : A(x', y^{(\ell)}) < \hat{q} \right\}, \tag{3}$$

that is, for each label $y^{(\ell)}$, we first calculate the corresponding score $A(x', y^{(\ell)})$, then if the score is less than $\hat{q}$ we include the label in Γ^α, the set of predicted labels.

CP guarantees that the probability that the true label is contained in Γ^α is approximately $1-\alpha$; this is known as "marginal coverage." More concretely, it can be shown [45] that

$$1 - \alpha \leq \Pr(y^{(\hat{\ell})} \in \Gamma^\alpha) \leq 1 - \alpha + \frac{1}{N+1}, \tag{4}$$

such that if N, the number of calibration data points, is sufficiently large, we recover the standard approximate result of $1-\alpha$.

Is this useful? Practitioners in the field will no doubt know that there is a well-built-up statistical literature on decision theory behind the FAR and $p_{\rm astro}$ introduced in Sec. II (and we will explore this in detail in our toy model; cf. Sec. IV). However, as discussed, pipelines can be miscalibrated and disagree with one another. The core motivation behind studying *CP* is that we can treat the statistical quantities arising from pipelines as heuristics and use the calibration dataset to adjust it, ensuring robust performance. As we will see later in Sec. VI C: this calibration process can, in fact, be viewed as a generalization of the empirical measurement of the FAR itself.

It is worthwhile to consider how *CP* quantifies uncertainty in the label. As scientists, we are used to talking about uncertainty on a measurement, e.g., a real-valued number accompanied by an uncertainty interval. *CP* can also tackle this problem (the realm of parameter estimation or regression), but in our current context, we do not have a real-valued number; instead, we have a label. For example, should we classify this chunk of data as containing a "signal" or just "noise"? *CP* provides uncertainty on the point prediction made by an underlying classifier by introducing the prediction set Γ^α. Inspecting Eq. (3), one can see that, for binary classification of signal or noise, the four possible prediction sets are the empty set $\varnothing$, one of two singleton sets {noise} and {signal}, or the double label {noise, signal}. As an anthropomorphic explanation, when asked "do these data contain a signal or noise?" the *CP* algorithm can respond "neither," "noise," "signal," or "either noise or signal."

Varying the error rate for a fixed test data point will vary the size of the prediction set. In the extremes, α close to zero or one, the *CP* algorithm will be forced to respond

with the double label or empty set (in the case of binary classification). Between the extremes, the performance will depend on the problem setup and choice of nonconformity score (we will demonstrate this later). This observation leads to the identification of what is known as the *CP* "confidence" [46], which we discuss later in Sec. IV C.

IV. CONFORMAL PREDICTION FOR A TOY COSMIC-RAY DETECTOR

We now provide a guide to *CP* in the context of classification and a simple astrophysics problem: a cosmic-ray detector. We will describe the problem and implementation qualitatively here, but the reader may wish to refer to the data release associated with this article, which contains program code to reproduce all parts of this section [47].

A. Problem setup

Consider a toy cosmic-ray detector consisting of a Geiger counter, which records the number of incidents of ionizing radiation it receives per minute while pointing to the sky (this example is not intended to be realistic but indicative of typical astronomy problems). Absent a cosmic ray, the detector will be subject to background radiation from terrestrial sources, which we model as Poisson distributed with a mean of λ_b counts per minute. The detector will observe a cosmic ray as a transient burst of N_c ionizing particles in some time δt, which, for the sake of this discussion, we take to be $\delta t \ll 1$ min. As such, we can identify and localize a cosmic ray in the data by searching for minute-long bins where the count rate exceeds the background. The excess amount will depend on N_c, which we will model again as Poisson distributed with mean λ_c. Finally, we will also model the number of cosmic rays as Poisson distributed with some rate λ_r per minute. In Fig. 4, we provide an illustrative example of data from our toy detector showing minute-long bins with background, clear cosmic-ray events (far above the background), and marginal cases in between.

The standard statistical search algorithm used in cases such as this to identify if a bin contains a cosmic-ray event is the frequentist one-sided p-value or, equivalently, the FAR. Namely, for an observed count c' and given the background rate λ_b,

$$\mathrm{FAR} = \frac{1}{T}\Pr(c \geq c'|\lambda_b) = \sum_{c=c'}^{\infty} \frac{\lambda_b^c e^{-\lambda_b}}{c!}, \tag{5}$$

where T is the bin duration of 1 minute. Note, for this toy model, we know the FAR in closed form; this differs from the empirical FAR, Eq. (1), we use in gravitational-wave astronomy.

Finally, our search algorithm proceeds by applying a threshold to the p-value or FAR: bins above the threshold likely contain a cosmic ray, while those below do not. In Fig. 4, we apply a p-value threshold of $1/20$ or, equivalently, a FAR of 1 per 20 min. At this threshold, we can identify four categories: several actual signals are identified (true positives, TP), but four background events above the threshold are identified as cosmic rays (false positives, FP). Meanwhile, several cosmic rays are missed and classified as background (false negative, FN), but most background events are correctly classified as background (true negative, TN). The nonzero counts of FP and FN are not a deficiency of the algorithm but rather inherent: with the true labels colored in Fig. 4, it is obvious which contains a cosmic ray and which does not, but our search algorithm has only the count rate leading, inevitably, to errors in classification.

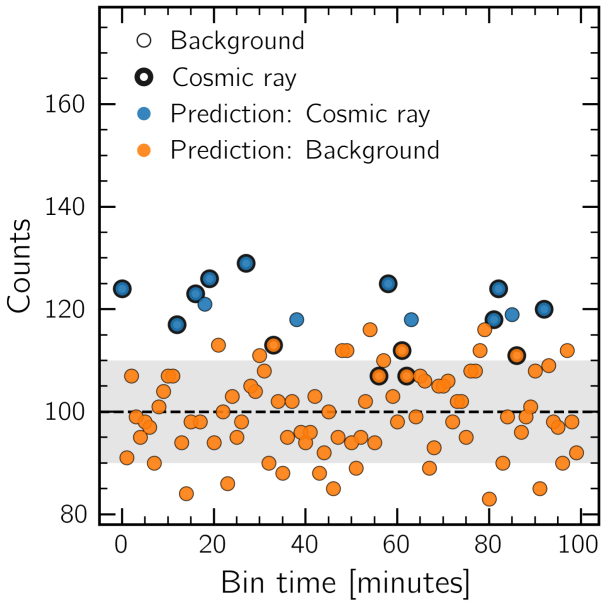

FIG. 4. An illustrative example of data from our toy cosmic-ray detector. Each data point records the number of counts within a minute-long interval or bin. Thick circles mark bins containing a cosmic ray. Data points are filled according to the prediction of the FAR detection approach: blue circles correspond to data points which surpass the threshold and, hence, where we reject the null hypothesis. In contrast, orange circles indicate those that are consistent with background noise.

Of course, this is a well-studied problem of statistical decision theory (see, e.g., Cowan [48]). In Figs. 5 and 6, we reproduce two standard figures of merit which demonstrate this behavior. First, the receiver operating characteristic (ROC) curve shows the true positive rate against the false positive rate. The ROC curve is generated by varying the FAR threshold, repeatedly simulating our cosmic-ray detector, and empirically measuring the two rates. The curve demonstrates the trade-off between true positives and false positives possible with our given search algorithm: points closer to the ideal case (top-left corner) are better in maximizing the true positive rate while minimizing the

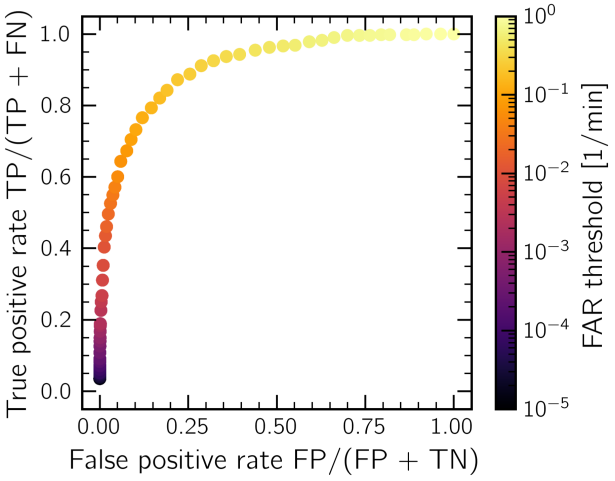

FIG. 5. Measured ROC curve for the simple cosmic-ray detector search algorithm. We measure false positive and true positive rates while varying the FAR (or, equivalently, the p-value) threshold.

false positive rate. Second, in Fig. 6, we show an alternative visualization of the same data: the precision and miss rate. Considering the case of a catalog of gravitational-wave signals, these are of more direct relevance. The precision tells of the purity of the catalog. If the precision is sufficiently close to 1, one can be reasonably assured the catalog is pure and does not contain any potentially biasing terrestrial artifacts. However, such a guarantee comes at a cost: the miss rate tends to 0 in the same limit, indicating the catalog size will shrink.

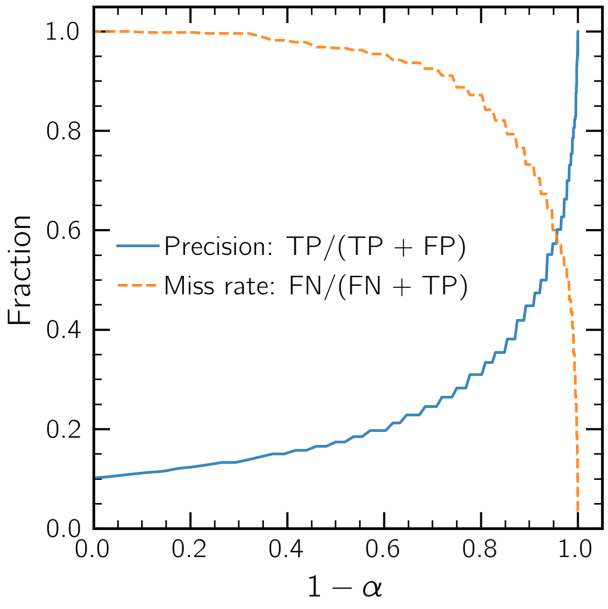

FIG. 6. The precision and miss rate for the simple cosmic-ray detector search algorithm as a function of $1-\alpha$ where α is the p-value (or, equivalently, the FAR).

B. Conformal prediction

At this point, we now step beyond the confines of classical statistical decision theory and introduce the application of *CP*. In this context, the cosmic-ray detector search algorithm described above can be considered a classification algorithm that produces a label $y \in \{\text{background}, \text{cosmic ray}\}$ (whereby "cosmic ray" we implicitly mean there is both a cosmic ray and background).

We apply the *CP* approach defined in Sec. III to our cosmic-ray detector problem. We generate a large set of calibration data points consisting of simulated data and the true classification (i.e., whether a cosmic ray was present or not). Next, we define our nonconformity score. We choose to use the complement of the Poisson probability mass function (noting that for the background + cosmic-ray case, the sum of two Poisson distributed variables is itself Poisson distributed with a rate equal to the sum of the rates), i.e.,

$$A(x, \text{background}) = 1 - \text{Poisson}(x, \lambda_b), \tag{6}$$

$$A(x, \text{cosmic ray}) = 1 - \text{Poisson}(x, \lambda_b + \lambda_c). \tag{7}$$

In Fig. 7, we visualize our nonconformity scores, showing that close to the mean, the nonconformity is at a minimum for each class, while away from these, they are close to unity. We note that the absolute magnitude of the variation in nonconformity measure is not important: what matters is the relative quantile they appear when ranked by the conformal algorithm. In this sense, the relative magnitude between classes is important (though this will not be the case later when we consider the class-conditional Mondrian conformal prediction later on).

Once our nonconformity score is defined, we can apply the conformal algorithm to new test data given some choice

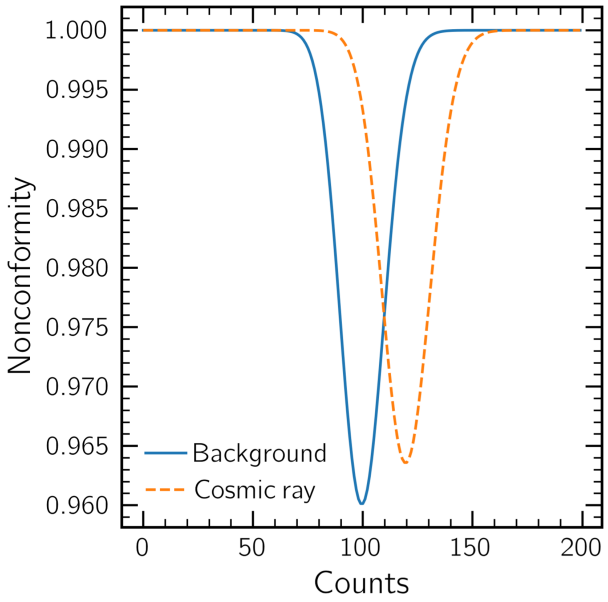

FIG. 7. Visualization of the nonconformity scores expressed in Eqs. (6) and (7).

of α. For each data point, the output of the algorithm will be the prediction set Γ^α. In our binary case, Γ^α can be the empty set $\varnothing$, one of two singleton sets {background} and {cosmic-ray}, or the double label {background, cosmic ray}.

The marginal coverage guarantee, Eq. (3), states that, if implemented correctly, the correct label will be in Γ^α a fraction $\sim 1-\alpha$ of the time. To check this, in Fig. 8, we plot the empirically measured coverage after applying the conformal algorithm to a large simulated cosmic-ray dataset. The marginal coverage (the number of times the true label appears in the prediction set) follows the one-to-one mapping guaranteed by Eq. (3), demonstrating proper algorithm implementation. There is some variation when $1-\alpha$ is close to 0 as the set sizes become small; moreover, the steplike nature of the empirical coverage arises from the discrete nature of the Poisson data in our toy model.

Figure 8 also provides an insight into the limitation of the simple *CP* algorithm: the coverage guarantee applies only to the marginal, not the conditional labels. As a result, the conditional labels may be over- or undercovered (i.e., exceed the allowed error rate). We see this manifest in Fig. 8 for the cosmic-ray label, which strays away from the diagonal. This is problematic: in gravitational-wave astronomy, we are not interested in ensuring that the label is correct as averaged over both the signal and noise labels. We want the validity guarantee [i.e., Eq. (3)] to apply to conditional labels. To achieve the guarantee for all labels individually, we can use Mondrian conformal prediction (MCP) [49], where the data are split by class, and then the conformal prediction

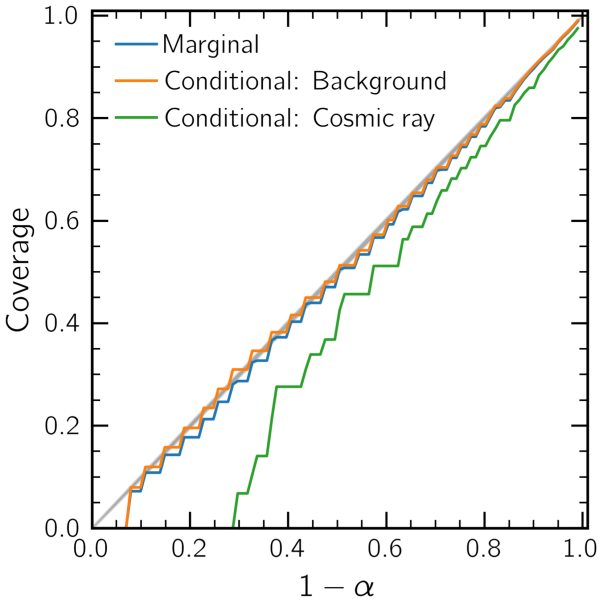

FIG. 8. The empirically measured coverage (the fraction of events for which the true label is in the prediction set) for the cosmic-ray test dataset after applying *CP*. A gray band marks the 95% binomial confidence interval expected given the size of the test data; we see variations around this due to the discrete nature of the underlying data.

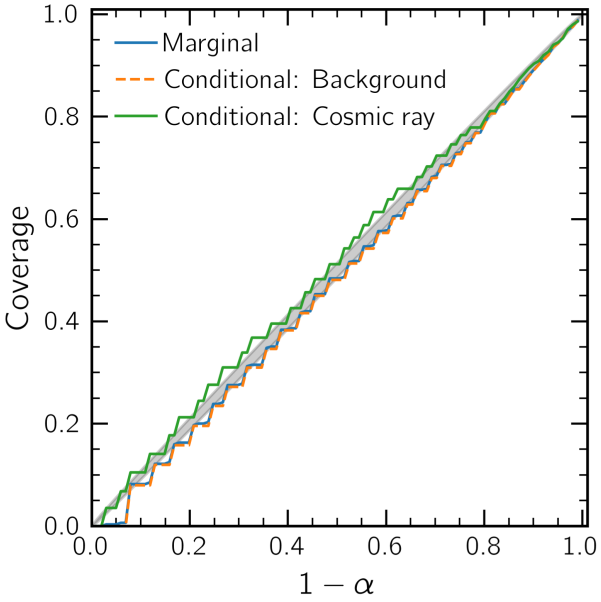

FIG. 9. The empirically measured coverage (the fraction of events for which the true label is in the prediction set) for the cosmic-ray test dataset after applying MCP. A gray band marks the 95% binomial confidence interval expected given the size of the cosmic-ray test data; we see variations around this due to the discrete nature of the underlying data.

algorithm is applied to each group separately. Using this technique, both the conditional labels are guaranteed to follow Eq. (3) and, by extension, the marginal labels do too.

The cost of MCP is that the number of calibration data points entering Eq. (3) is no longer the total number but the number per label. Therefore, the intrinsic error on rare classes consistently exceeds more common labels by design. We apply the simple classwise algorithm where the possible labels define the groups [49]. However, more advanced approaches are possible: see Ding *et al.* [50] for a formal introduction to the topic and discussion of a clustered algorithm capable of extending to many sets.

To apply MCP, we split our calibration dataset into simulated data points containing a cosmic ray and those that do not. Then, we apply *CP* to each label and the corresponding calibration set separately for the test data. For this reason, unlike the standard *CP* algorithm, the relative values between nonconformity measures do not matter in MCP. In Fig. 9, we reproduce Fig. 8 but having applied MCP. Now, Eq. (3) is valid for both the marginal and class-conditional labels. Finally, in Fig. 10 we provide an illustration of the example data from Fig. 4, but demonstrating the application of MCP; this illustrates how CP adds uncertainty to the prediction near to the boundary where the distributions overlap.

C. Confidence

There is a defined quantity within the *CP* framework known as the confidence [46]. This arises from noting that

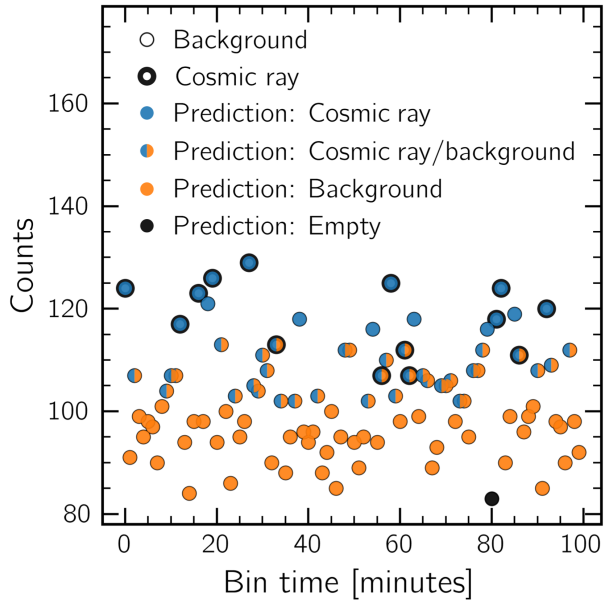

FIG. 10. The illustrative example of data from Fig. 4, but with the labels as predicted by the MCP algorithm and using $\alpha = 0.1$ (i.e., at a 90% coverage guarantee).

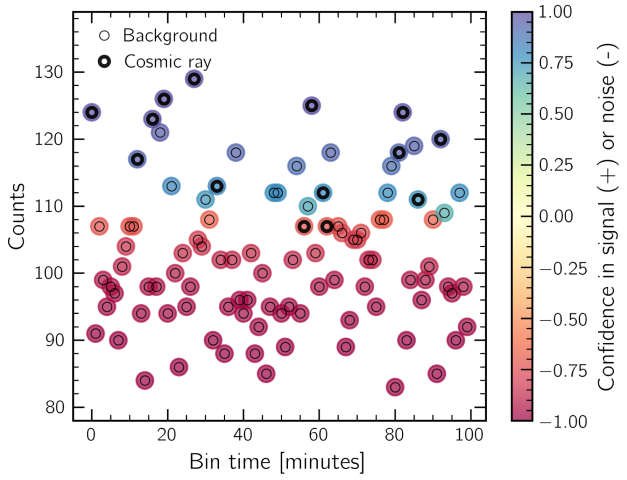

FIG. 11. The illustrative example of data from Fig. 4 colored by the confidence as defined in Definition 1. To aid visualization, positive values are assigned to data points where the single-label prediction is cosmic ray while we assign negative confidences to those where the single-label prediction is background (i.e., values closer to -1 indicate greater confidence in the noise label).

the Γ^α prediction sets are nested, such that if $\alpha_1 \geq \alpha_2$, then $\Gamma^{\alpha_1} \subseteq \Gamma^{\alpha_2}$. Since the size of Γ^α is a discrete quantity, it varies in steps, and these change points can be used to assign significance statements. This observation leads us to the standard definition of confidence:

Definition 1. The confidence is the value of α such that the size of Γ^α changes from 1 to 2 (i.e., the point where we go from the single to the double label).

Necessarily, each data point has a unique confidence assigned to whichever label is the *single* label given the data.

In Fig. 11, we take our demonstration cosmic-ray data and add the confidence, assigning [0, 1] as the confidence for data points with single-label prediction cosmic ray and flip the confidence to $[-1, 0]$ for data points with single-label prediction "background" (this is nonstandard, but allows in the binary case to plot the confidence on a single diverging color scale). From this figure, we observe a sharp divide near the boundary between the nonconformity scores of the two labels (cf. Fig. 7). This notion of confidence does have uses; for example, it automatically produces a potential decision algorithm for calling something a signal: only those data points for which the single label is cosmic ray. However, it is limited in that it does not allow one to talk about the confidence that an arbitrary data point contains a signal because, for those with a single-label background, the confidence is the background confidence.

To further understand the confidence, we note that, in this toy example, it is a function only of the observed count rate. Therefore, as in Fig. 12, we can plot the confidence as a function of the count rate to see the mapping. In this figure, we see that, at a count rate of 110 (the point where the nonconformity scores of background and cosmic-ray labels are equal, cf. Fig. 7), the confidence flips between the cosmic-ray and background single label. There is a minimum, and on either side, the confidence monotonically increases for either label.

This motivates us to consider an alternative definition, the conditional confidence:

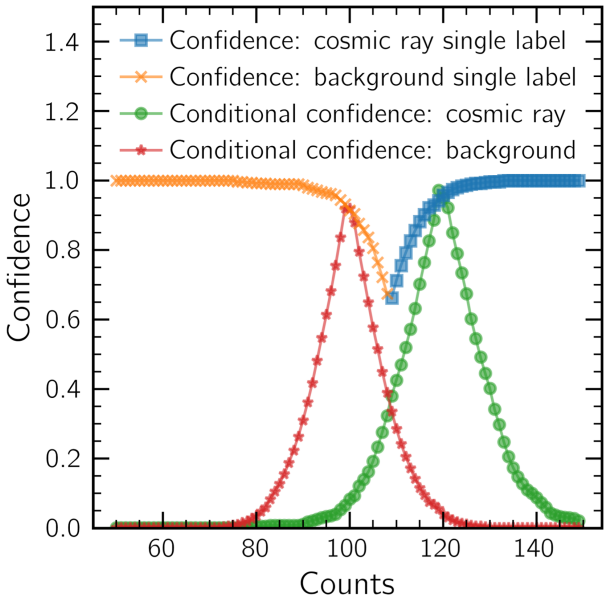

FIG. 12. The mapping from counts to confidence (cf. Definition 1). In blue, we show the confidence of counts where the single-label prediction is cosmic ray, in orange cases where the single-label prediction is background. We also plot the mapping to the conditional confidence (cf. Definition 2), the cosmic-ray (green) and background (red) labels.

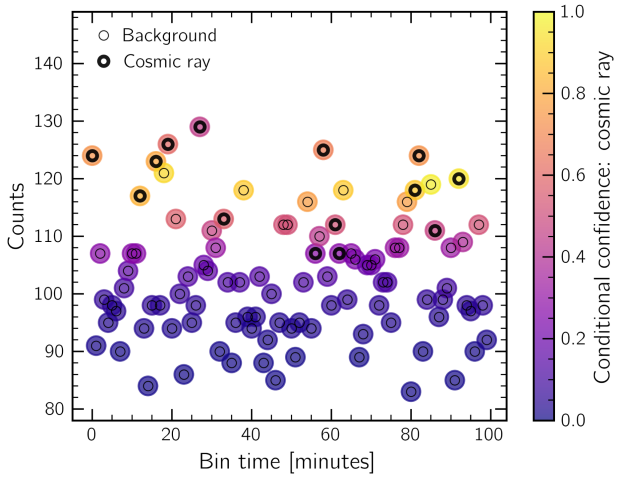

FIG. 13. The illustrative example of data from Fig. 4 colored by the "conditional confidence" (i.e., the minimum value of α such that the conditional label is included in the set, cf. Definition 2) for the cosmic-ray label.

Definition 2. The conditional confidence in label y is the minimum value of α such that $y \in \Gamma^{\alpha}$.

We add this to Fig. 12 for both the cosmic-ray and background labels, demonstrating that it can be calculated for any data point. Comparing Figs. 12 and 7, it is apparent that, in this example, the conditional confidence is the scaled complement of the nonconformity score. In a sense, this may seem circular. However, it is worth noting that the conditional confidence depends on the distribution of nonconformity scores in the calibration set and not solely on the nonconformity score itself. Intuitively, the conditional confidence in label y can be understood as the probability (interpreted as a relative frequency) that the true label is y as measured from the calibration dataset. We believe conditional confidence is useful in providing an intuitive guide to understanding the significance associated with each label for a given data point. To conclude, we finally apply the conditional confidence to our demonstration data in Fig. 13 which, contrasted with Fig. 11, demonstrates a smoother variation in assigned confidence and the ability to assign confidence in the cosmic-ray label to all data points.

D. Measuring performance by set size

Figure 9 may give the impression that we achieved perfect performance at no cost: the calibrated *CP* label sets always contain the true labels a fraction $1-\alpha$ of the time despite us never testing the performance of the conformity scores. However, we did not consider the *set size*, i.e., how many labels are given singleton labels cosmic ray or background, the double label, or no label at all? Indeed, the set size is critical to practical utility and where we should measure the performance of our nonconformity scores.

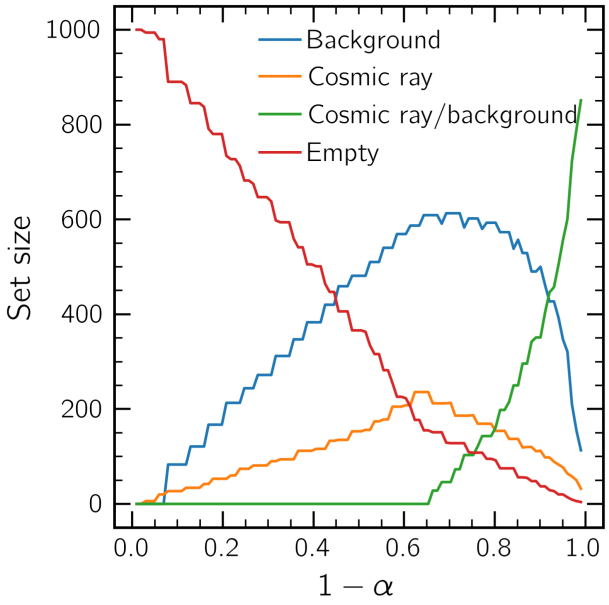

FIG. 14. The set sizes for the four possible prediction sets after applying MCP to 1000 test points for the cosmic-ray detector problem.

In Fig. 14, we plot the set size for all four possible prediction sets as a function of $1-\alpha$. In doing so, we show the performance: the ability to identify cosmic-ray and background events uniquely varies as a function of the allowed error rate. At the lower extreme, we have the limiting behavior of the algorithm. Namely, for $1-\alpha \sim 0$ (the maximum allowed error rate), all data points are in the empty set while the size of the singleton and double labels is close to 0. For $1-\alpha \lesssim 0.6$, the set size of the singleton labels grows linearly with the size of the empty set decreasing. Above $1-\alpha \sim 0.6$, the set size of the singletons and empty set decrease while the set size of the double label rapidly increases.

Figure 14 explains why there is no free lunch with *CP*. While we can choose $1-\alpha$ arbitrarily close to 1 (i.e., minimize the allowed error rate), this comes at the cost of increasing the size of the double label. That is, the cost is a majority of triggers for which the algorithm is essentially uninformative. Here, there is a parallel with Fig. 6 in which we saw that choosing a conservative threshold increased the precision at the cost of increasing the miss rate. Such behavior is unavoidable, but by measuring the set size, one can compare and optimize choices of nonconformity score.

E. Performance of a poor nonconformity score

Finally, it is helpful to take an illustrative example of what happens when the nonconformity score performs poorly. To demonstrate this, we take our cosmic-ray detector example and consider an alternative choice of nonconformity score,

$$A(x, \text{background}) = 1 - \text{Poisson}(x, \lambda_b), \tag{8}$$

$$A(x, \text{cosmic ray}) = U(0, 1), \tag{9}$$

i.e., while the background score stays the same, we replace the cosmic-ray score with a uniform random number generator. We show the results by applying this to our demonstration data in Fig. 15. At first, it may appear to still perform reasonably well: most of the cosmic-ray events are labeled as cosmic ray. However, on closer inspection, we see that almost all the noise events are given the double label, multiple prominent cosmic rays have no label assigned, and background data points are labeled as cosmic rays.

This choice of the nonconformity score is extreme but yields insights into what to expect if a poor choice is made for the nonconformity score. We can further study the behavior by looking at the set sizes as a function of $1-\alpha$; this is done in Fig. 15 and shows that at $1-\alpha = 0.5$, labels are randomly assigned between the four choices while at either extreme either no label is assigned or the double label.

The set size is one way to measure the performance of a nonconformity score. For example, comparing Figs. 14 and 16 we see that, around $1-\alpha \sim 0.7$, the standard nonconformity score produces more single labels than either the double or background. Meanwhile, this is never true for the alternative [i.e., Eqs. (8) and (9) which are intentionally broken] nonconformity scores demonstrating that the informative nonconformity measure outperformed the alternative. The choice of nonconformity score can therefore be viewed as an optimization problem. However, the choice of objective function is itself subjective and will depend on the use case. For example, one option is to choose a nonconformity score that minimizes the number of double labels, aiming to increase the algorithms' capacity to unambiguously label the data. However, such a choice may come at the cost of increasing the empty label set. Alternatively, one may choose to maximize the TP rate (or minimize the FP rate) at some fixed α. Extending this idea, the nonconformity score itself can be parametrized, enabling direct optimization (see, e.g., Colombo [51]). Regardless of the methodology, the choice of objective function for the optimization will always be subjective and the best choice will depend on the overarching use case. For gravitational-wave astronomy, we anticipate some combination of maximizing the number of single labels while minimizing the number of false positives, but we intend to explore this in future work.

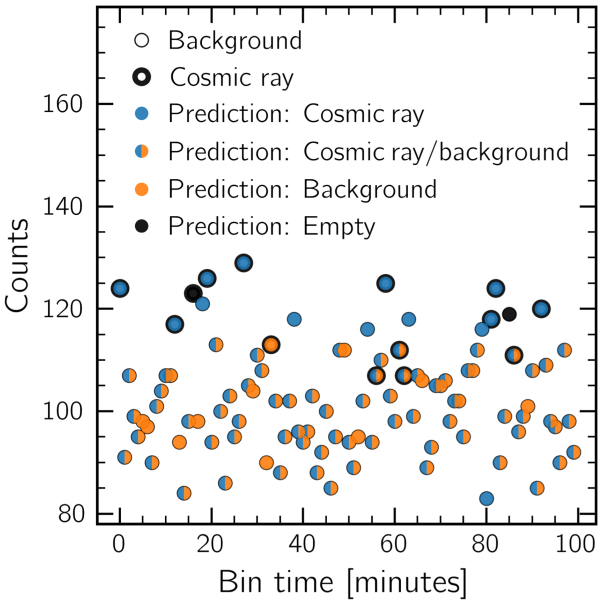

FIG. 15. Reanalysis of Fig. 10 with $\alpha = 0.1$, using Eqs. (8) and (9): a noninformative conformity measure for the cosmic-ray label.

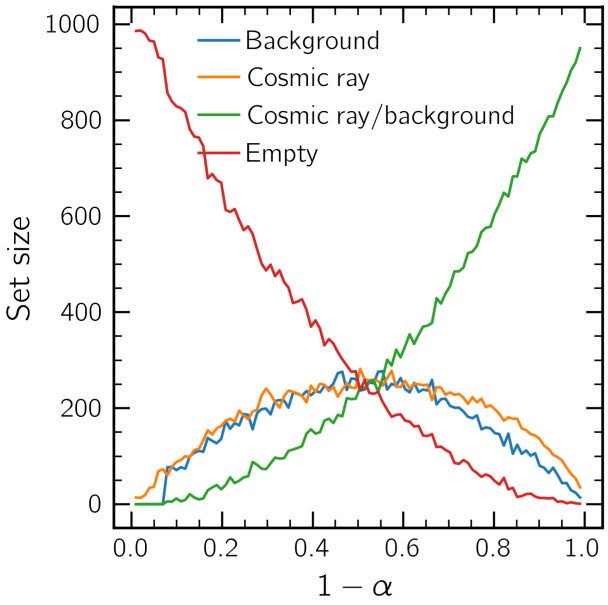

FIG. 16. The set sizes of the four possible prediction sets after applying MCP to 1000 test points with the nonconformity scores given in Eqs. (8) and (9).

V. CONCLUSION: TOY MODEL

In this section, we have used a simplistic toy model to introduce *CP*. In the main, we use this as a tool to understand *CP* and not as a demonstration of the application of *CP* to realistic astrophysical problems. We recognize that there are steps that do not transcend, e.g., here, we know the statistical properties of the signal and noise distributions perfectly and can use these to construct a nonconformity score. Nevertheless, we hope it may prove useful as a starting point for others to apply *CP* using the accompanying notebook [47].

VI. CONFORMAL PREDICTION FOR GRAVITATIONAL-WAVE ASTRONOMY

Having introduced *CP* for a simple toy model, we now extend the discussion to gravitational-wave astronomy. We will focus on the use case of modeled transient searches

for CBC signals. However, the discussion applies generally since the standard statistical framework is applied across the field.

Our primary task is to define the nonconformity measure $A(x, y)$. Considering the binary classification problem, signal or noise, two obvious initial choices exist: using the FAR or the Bayesian p_{astro} quantities. For source classification, e.g., binary black hole, neutron star black hole, binary neutron star, or terrestrial, one could use the multiclass *CP* algorithm and the Bayesian probabilities provided by the pipeline for each source class. Therefore, these choices are readily applied to the outputs of existing pipelines, which is what we choose to do in this work.

However, *CP* offers scope for further development. For example, the FAR used by pipelines uses a ranking statistic combining the matched-filter SNR and χ^2 statistic amongst other quantities. Such a combined ranking statistic can itself be used as a nonconformity score: in effect, the "calibration" dataset of *CP* is then analogous to the background data used in a traditional search pipeline. Building on this idea, if the combination is parametrized, one could optimize the ranking statistic (nonconformity score) to minimize the counts of the empty set of multilabel prediction sets on some test data. Such an idea builds on a similar application by McIsaac and Harry [52], which seeks to maximize the separation of signals and noise. Many more such innovations are likely possible.

A. Using conformal prediction to calibrate multiple competing pipelines

To demonstrate the application of *CP* to gravitational-wave astronomy, we will use the results of a recent mock data challenge (MDC) study in advance of the LVK fourth observing run [33]. In this MDC, four low-latency CBC online search algorithms were applied to a real-time data replay from the third observing run. Simulated signals were added to the data at a rate much greater than the anticipated astrophysical rate under current detector sensitivities. This higher rate was used to stress test the low-latency infrastructure: the primary goal of the MDC was to measure expected performance in producing public alerts used to trigger event follow-up. Taking the MDC data, we adjust classifications for all real gravitational-wave detector events present in the MDC, but do note there are potential subthreshold signals that remain. We also remove all early warning triggers from the MDC and use the corrected p_{astro} values from Ray *et al.* [43].

The MDC data products provide a perfect test bed for *CP*. The increased rate produces a sizable set of simulated triggers, e.g., points in the data stream that the search pipelines identify as likely to contain a signal. Most recorded triggers in the MDC are simulated signals (this differs from the astrophysical scenario where, at a high FAR threshold, most triggers will be nonastrophysical noise). Moreover, the configuration of the pipelines was in development during the MDC, leading to imperfect performance. For these reasons, the performance of the pipelines is not representative of the tuned performance expected during the run. This point is discussed within Chaudhary *et al.* [33] specifically for the case of PyCBC: "The FAR values for injections recovered during the MDC are subject to a substantial upward bias due to the high rate of high-SNR injected events, which significantly influences the background estimation." As a result, in the context of candidate significance estimation, we can consider the MDC data as the application of poorly calibrated pipelines to a given dataset. It, therefore, is a good test bed to show how *CP* can automatically calibrate the pipelines. However, we stress that the following discussion should not be taken as indicative of the performance of the pipelines, only as an example where they are known to be ill tuned.

Let us begin by studying the performance of the pipelines using traditional significant estimation approaches. We start by thinking about the catalog of events that would be produced at a given threshold. In Fig. 17, we plot the purity of the resulting catalog as a function of p_{astro}; we present results separated by pipeline. We calculate the purity as the fraction of triggers with p_{astro} greater than the threshold which pertains to an injected signal. We plot the actual purity (the true number of simulated signals in the trigger set) and the estimated purity: the sum of the p_{astro} for all triggers above the threshold. The sum of p_{astro} to estimate the number of astrophysical signals is

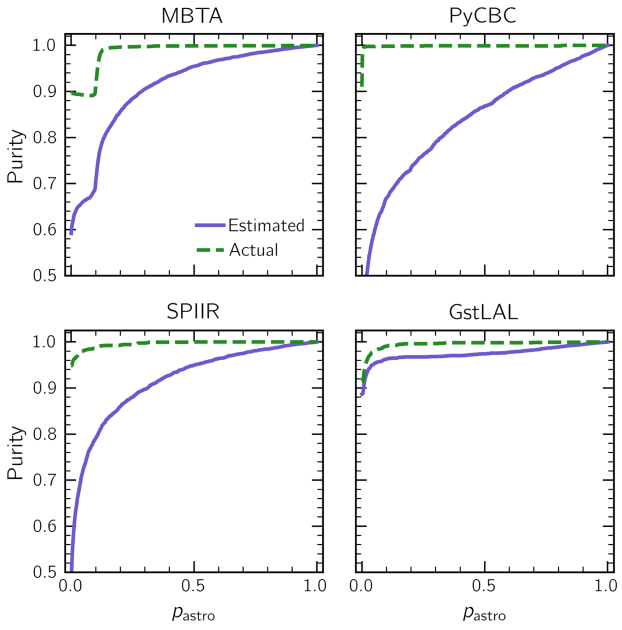

FIG. 17. The estimated and actual purity for the MDC results as a function of the p_{astro} threshold split by pipeline. Estimated purity refers to the sum of p_{astro} above the threshold, while actual purity refers to the count of triggers pertaining to signals above the threshold. We use purity here as it is the common language of the field. However, we note that it is identical to the coverage defined in the field of *CP*.

commonly used in the context of a catalog of triggers (see, e.g., Abbott *et al.* [5]). It formally amounts to the posterior-estimated number of foreground events in the Farr *et al.* [39] framework. Figure 17 shows varying behavior by pipeline, with all pipelines underestimating the actual purity by varying amounts. (We note that, due to the presence of potential subthreshold real signals in the MDC data, the "actual" estimate here is potentially biased; however, given the expected purity of subthreshold candidates in GWTC-3 [5], the level of bias is at most a few percent.) By comparison, the advantage of *CP* is that α, the allowed error rate of the algorithm, maps directly onto the actual purity of the resulting catalog.

To demonstrate *CP* in practice, for the set of candidates from each pipeline, we evenly split the MDC data results into a calibration and test set. We then apply MCP using the FAR as the nonconformity score for signal and the inverse False Alarm Rate (iFAR) as the nonconformity score for noise. This way, we use the pipeline outputs directly without adding additional information. We then apply MCP to each trigger in the test dataset, using the calibration data for producing a prediction set. Note that the computational effort required for this step is negligible (a few CPU seconds on any modern computer).

In Fig. 18, we plot the label coverage for each pipeline, demonstrating it satisfies Eq. (3), i.e., for all α, the fraction

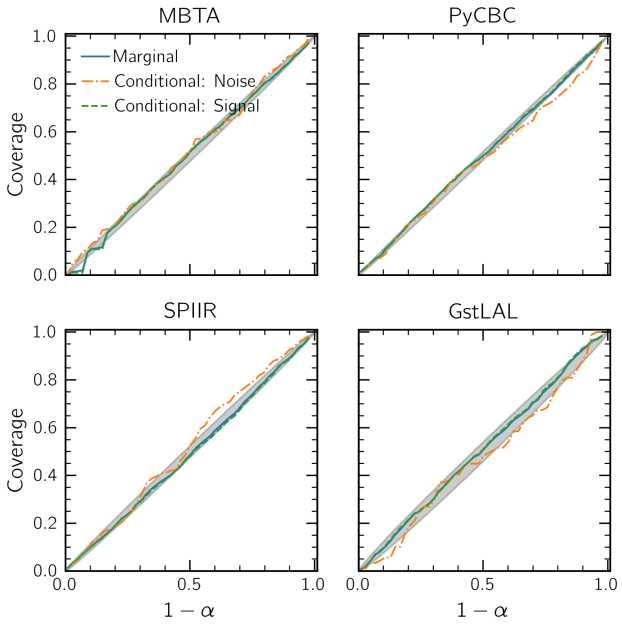

FIG. 18. The marginal and conditional coverage for all MDC results after applying MCP, demonstrating they satisfy the validity guarantee. Recall that the marginal coverage is averaged over all labels, while conditional coverage is as applied to a single label at a time. A gray band marks the 95% binomial confidence interval expected, given the size of the entire test data for each pipeline. Note that for the conditional labels, the size of the effective test dataset is smaller, and therefore, the anticipated Poisson counting error can be larger, as is the case of the GstLAL conditional noise label.

of test triggers which contain a simulated signal has a one-to-one correspondence with $1-\alpha$. Moreover, we note that all pipelines satisfy this: irrespective of their underlying performance, once calibrated by *CP* the coverage guarantee is ensured. We now note that what is known in the field of *CP* as coverage is equivalent to the catalog purity. As such, Figs. 18 and 17 can be contrasted to show how calibrating with *CP* regularizes the meaning of the threshold between pipelines. The implication is that once calibrated by *CP*, the catalog produced at a fixed α threshold contains an *a priori* known contamination rate: α. Therefore, downstream analysis can decide the contamination rate they are willing to accept and then use that to set the threshold for inclusion.

B. Understanding individual events: Confidence

In the last subsection, we saw how a catalog could be created by applying MCP to calibrate the significance estimates. Such an application guarantees the purity of the resulting catalog. It is, therefore, directly applicable to the case of population analyses, where one often needs to control the purity over a set of triggers. However, this leaves the question of assessing individual events and deciding if they are astrophysical, which we now discuss.

In the traditional framework, candidate significance is assessed by combining the FAR, $p_{\rm astro}$, their constituent elements (e.g., the χ^2 statistic), and a deep knowledge of the performance of the pipeline. For example, the first direct observation of gravitational waves from GW150914 [4] reported a FAR of 1 event per 203,000 years (and gave an equivalent $>5\sigma$ estimate). However, once a source class is established, $p_{\rm astro}$ is generally the preferred mechanism to identify new events (for example, independent reanalyses use this criterion Venumadhav *et al.* [53]). However, for newly detected source classes, because $p_{\rm astro}$ requires an astrophysical model of the rates, which is generally poorly constrained, it is common to revert to a more detailed study of the FAR (see, e.g., the discovery of the first neutron star black hole mergers [54]).

In the *CP* framework, we can use the confidence to assess candidate significance. As discussed in Sec. IV C, one can compute either the standard definition of confidence, Definition 1, or the conditional confidence, Definition 2. We now consider how these definitions of the confidence can be applied to CBC signals using the MDC for illustration.

In the left-hand panel of Fig. 19, we plot the standard confidence (Definition 1) for all triggers in the MDC against their iFAR. We find a one-to-one mapping, which is expected since we use the FAR as the nonconformity score for the signal label. The standard confidence that the data contain a signal can only be computed when the single-label prediction is for a signal (see Sec. IV C). Therefore, we find there is a minimum iFAR below which the conditional confidence that the data contain a signal

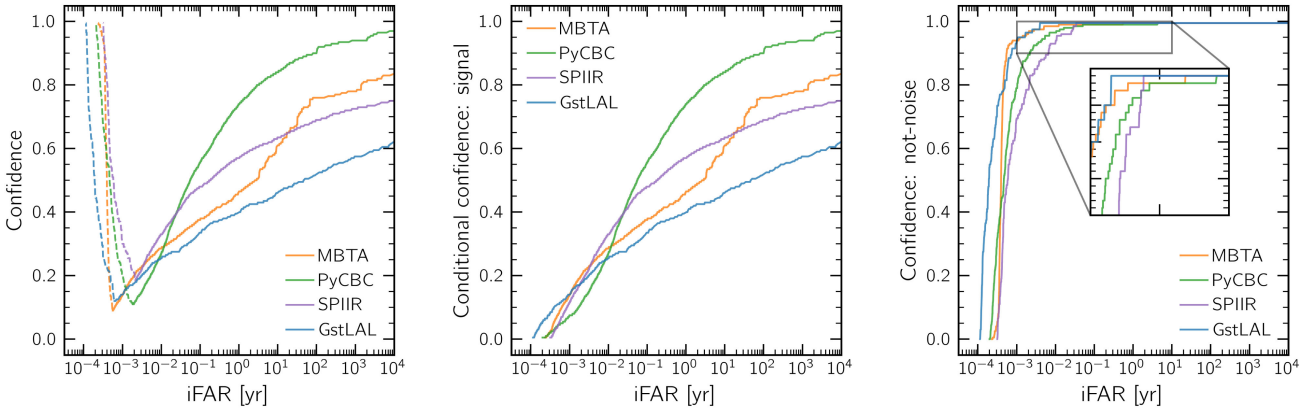

FIG. 19. The relation between the iFAR and three definitions of confidence within the *CP* framework: the standard confidence given in Definition 1 (left-hand panel), the conditional signal confidence given in Definition 2 (middle panel), and the not-noise confidence given in Definition 3 (right-hand panel). For each definition, we plot the confidence against the iFAR for all triggers (separated by pipeline) in the MDC. For the standard definition, Definition 1, the confidence that the data contain a signal can only be calculated when the single-label prediction is for a signal (see Sec. IV C); we mark these points by a solid line in the left-hand panel. Meanwhile, for values of the iFAR where the single-label prediction is for noise, we use a dashed line. We, therefore, see a turnover in the leftmost panel, a minimum iFAR below which we cannot assign any confidence that the data contain a signal. We sort triggers by iFAR to produce a continuous line showing the learned mapping. In all cases, we truncate the figure at an iFAR of 10^4 yr for visualization purposes: the mapping extends up to the maximum iFAR in the dataset and monotonically approaches unity in that limit. In the right-hand panel, we add an inset showing the behavior as each curve approaches unity.

cannot be computed. Instead, we can compute the confidence that the data are noise (since, in this binary case, that is now the single-label prediction). We illustrate this by adding the noise confidence as a dashed line.

In the middle panel, we go on to show the mapping between the conditional confidence in the signal label, Definition 2, against the iFAR. Unlike the standard confidence, the conditional confidence can be computed for all values.

For both the standard and conditional confidence, we note that they behave broadly as we expect: the confidence increases monotonically with the iFAR. However, it is notable that the mapping is at odds with the expectation of seasoned analysts in this field: namely, we find that even at a FAR of 1 per 1000 years, the confidence of some pipelines is barely above 0.5. For comparison, in Fig. 3, at a FAR of 1 per 1000 years, all pipelines report a p_{astro} close to unity.

Moreover, the confidence is pipeline dependent, with substantial disagreements between pipelines. This occurs due to our choice of nonconformity score: we use the FAR. The nonconformity score ranks how signal-like the data are compared to the most significant signal in the data: smaller FARs are more signal-like. As a result, pipelines that have a long tail in the iFAR for signals will consequently produce less confidence at the same iFAR relative to pipelines with shorter tails. (It should be remembered, however, at this point that *CP* is distribution-free in the sense that the distributions are never explicit but learned via the calibration dataset.) There is nothing inherently wrong here, but we do concur that what is known as confidence in *CP* does not reflect what a gravitational-wave analyst might understand the term to mean.

If we would like the confidence to better reflect our understanding, we can either look at the choice of nonconformity score or the definition of the confidence. An obvious alternative choice for the nonconformity score is p_{astro}: however, since this is closely related to the FAR (cf. Fig. 3), we encounter similar issues. Meanwhile, it is worthwhile reflecting on why the seasoned analysts' intuition suggests that a signal with an iFAR of 1000 years should confidently be called a signal. This is because, if the pipeline is well calibrated (which we anticipate to be the case most often), then the iFAR intrinsically suggests the data are not consistent with the background. With this in mind, we define another definition of confidence, the "not-noise" confidence:

Definition 3. The not-noise confidence is the minimum $1-\alpha$ such that the noise label is not included in Γ^α.

Applying this definition in the right-hand panel of Fig. 19, we recover a mapping much more in line with expectation: we see a rapid increase in the not-noise confidence, and for values above 1 yr, the confidence is close to unity. This demonstrates the power of *CP*: it should be remembered that the underlying algorithm is distribution-free, it has learned this intuitive threshold directly from the calibration data. Moreover, if the underlying algorithm itself was not well calibrated, the confidence still would be (this would manifest as a significant departure from the four calibrated pipelines in the right-hand panel of Fig. 19).

The three definitions of confidence presented in Fig. 19 all offer different ways to assess the confidence we may

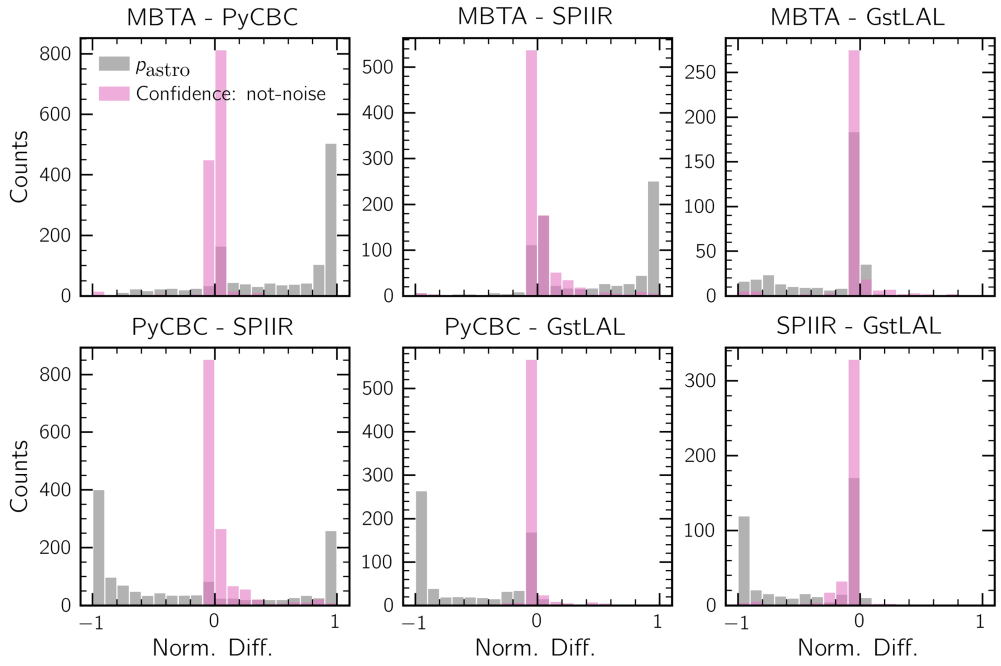

FIG. 20. Histogram of the normalized difference (i.e., the difference divided by the sum) in the not-noise confidence and p_{astro} for all pairs of pipelines in the MDC. Note, we filter to only cases where both pipelines identify the signal (defined as finding a trigger within a 0.1 s window) and take the closest match in trigger time. We also filter cases where p_{astro} is not predicted by one or both pipelines.

have in an individual event. However, we believe that further work needs to be done to identify which of these (or perhaps an alternative definition) is best suited to providing a summary of the significance of an individual event. Moreover, careful future study will need to be made of how these interact with the choice of nonconformity score. We also suggest that alternative choices of nonconformity be explored to see if these can better represent our understanding.

Finally, if the *CP* calibration has succeeded, we should expect it to regularize pipeline behavior, i.e., we would expect that the same event found by different pipelines would have a similar confidence. We would not expect it to give the same confidence to a given event since pipeline performance differs. To investigate this, in Fig. 20, we plot histograms of the normalized difference between the not-noise confidence for all pairs of pipelines. We also show the difference between p_{astro} for the same pairs. Notably, while the p_{astro} difference has a bimodal structure, with frequent cases in which the pipelines completely disagree about a candidate, the confidence difference peaks at zero, demonstrating a spread up to the extremes. This demonstrates that the confidence measured by *CP* regularizes behavior between pipelines by learning from the calibration dataset.

C. Conformal prediction as a generalization of the traditional framework

To conclude our discussion, we finally discuss how the *CP* and traditional FAR thresholds are related. In the traditional framework, to determine if the data contain a signal, we calculate the FAR [cf. Eq. (1)] and then apply a threshold: FAR′. If the FAR is below the threshold, we reject the null hypothesis and determine it is likely a signal. We can, therefore, formulate this in the language of *CP* by saying that the prediction set of the traditional framework is

$$\{\text{signal}:\text{FAR} < \text{FAR}'\}. \tag{10}$$

Formally, this is incorrect as it falls into the "inverse fallacy" in that by rejecting the null hypothesis, we assume the data contain a signal. However, in practice, it is very often done. Meanwhile, in MCP, if the signal nonconformity measure is given by the FAR while the noise nonconformity by the iFAR, the prediction set is given by

$$\{\text{signal}:\text{FAR} < \hat{q}_s\} \cup \{\text{noise}:\text{iFAR} < \hat{q}_n\}, \tag{11}$$

where $\hat{q}_s$ and $\hat{q}_n$ are (effectively) the $1-\alpha$ quantile FAR and iFAR of the calibration dataset (cf. Sec. III).

Comparing Eqs. (10) and (11), we now see the following three connections between the two methods in the binary classification case where the FAR (or equivalently the p-value) is used as the nonconformity score. We use these to explain the differences and advantages of *CP*.

First, in the traditional framework, the threshold for determining if the data contain a signal is chosen by hand. In contrast, in the *CP* framework, the threshold is automatically decided by the algorithm and calibration dataset (i.e., $\hat{q}$ is determined by the user choice of α). Of course, if

the FAR is already well calibrated, the *CP* framework offers no advantage in this respect. However, if that is not the case, *CP* calibrates the pipeline automatically.

Second, *CP* extends the labeling: while in the traditional framework, one either learns the data are a signal or not, for *CP*, the prediction set can be used to assess significance. That is, at a fixed choice of α, the set may contain both signal and noise: this provides the user with a means to understand the inherent uncertainty, and a choice of definition can be applied to calculate a confidence in a given label.

Finally, we see that in *CP*, one does not fall foul of the inverse fallacy: the signal label arises naturally from the definition of the nonconformity score without assuming it is the negation of the noise label.

Taken together, we therefore argue that *CP* can be viewed as an extension of the traditional statistical framework.

VII. DISCUSSION AND APPLICATIONS OF CONFORMAL PREDICTION

Conformal prediction offers a generalization of the traditional framework for significance quantification in gravitational-wave astronomy. In this work, we aim to introduce and explore *CP* in the context of CBC searches: we do not seek to demonstrate real application yet and envision this for future work.

We now outline three ways where *CP* may enhance existing efforts.

First, the conditional confidence can provide a calibrated alternative to the p_{astro} and FAR in assessing the significance of single events. A motivating question we posed in the Introduction is how to answer questions such as "do these data contain an astrophysical signal?" The traditional framework answers this by comparing the FAR to a threshold or with the astrophysical probability p_{astro}. In contrast, *CP* offers the confidence: the key difference between these concepts is that the confidence does not rely on an explicit astrophysical model like the p_{astro} and is learned from the performance of the pipeline on calibration data. As shown in Fig. 20, this moderates the differences between pipelines, leading to a more stable estimate of the significance.

Second, CP provides a means to the purity of a catalog. With *CP* we can circumvent the problem of determining a threshold on the significance by instead only requiring the user to specify the error rate. Specifically, given the appropriate tools, a user could set an error rate of 1% and then take all events where the signal label is in the prediction set and be assured by Eq. (3) that at least 99% of the catalog are astrophysical signals (within the bounds of the exchangeability assumption). As shown in Fig. 18, this guarantees the user that the catalog contains a fixed contamination fraction.

Finally, *CP* offers a framework to develop a postprocessing search pipeline combining the outputs from multiple search pipelines. Specifically, in future work, we will develop a parametrized nonconformity score combining the outputs from multiple pipelines into a single metapipeline. This has the advantage that the between-pipeline behavior can be regularized using the test and calibration data and we can optimize the score leveraging parameter-space-dependent pipeline performance.

For any of these applications to be successful, the critical missing ingredient is a large-scale MDC, which accurately captures the actual pipeline performance on realistic data. The MDC used in this work used an unrealistically high event rate and, therefore, is inappropriate for application to astrophysical signals. Indeed, this underlines the primary limiting factor of *CP*: the assumption of exchangeability between the calibration and test data. Ensuring this in practice will not be easy. Unlike many ML use cases, we must simulate the calibration dataset for gravitational-wave applications since we do not have a ready training dataset. In the simulation, assumptions must be introduced, e.g., about the waveform models and the rate: assessing and validating these will be critical. Moreover, using data from past observing runs breaks exchangeability as the detector sensitivity changes dramatically (Moreover, since it changes during an observing run, this is also a concern). In conformal prediction, such nonexchangeability cases are known as "distribution drift" and can be accounted for by applying weighted conformal procedures [45]. Nevertheless, we expect this to be a challenge for any successful application.

We acknowledge that the direction of *CP* is in many respects orthogonal to the overall direction of the field where the p_{astro} approach has become dominant. However, we believe that in some cases, end users of the data products do not sufficiently understand the assumptions and caveats of the many p_{astro} methodologies to interpret them fully. While p_{astro} offers a valuable and powerful approach, *CP* offers an alternative in which the end user can, given existing open access to the data and software, calibrate the pipeline themselves, allowing *CP* to learn the uncertainty inherent in the underlying method. Moreover, we want to emphasize that, for either the p_{astro} or FAR (or, equivalently, p-value approach), if the underlying assumptions are met, *CP* cannot improve on them. That is, *CP* does not offer a mechanism to improve the sensitivity of well-calibrated searches. However, it does enable calibration without requiring an understanding of the internal models or making asymptotic assumptions.

Finally, in this work, we have discussed the potential application for CBC search. However, *CP* may also find utility in other areas of the field, such as the low-latency alert products attached to open public alerts, the search for continuous gravitational waves from rapidly rotating neutron stars, or the search for bursts of GWs from unknown sources.

The source program behind Sec. IV is openly available from the Zenodo repository [47].

ACKNOWLEDGMENTS

We want to thank Geoffrey Mo for their support in accessing and understanding the MDC data products, Will Farr for help with the estimated purity in the p_{astro} approach, Anarya Ray for support with the GstLAL p_{astro} values, and Sebastian Khan for comments on the manuscript. We also thank Michael Coughlin, Deep Chatterjee, Tito Dal Canton, Reed Essick, Shaon Ghosh, Sushant Sharma-Chaudhary, Max Trevor, and Andrew Toivonen for the development of the MDC results used in this work. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. Computing support was provided by the Oracle for Research program.

[1] J. Aasi *et al.* (LIGO Scientific Collaboration), Advanced LIGO, Classical Quantum Gravity **32**, 074001 (2015).
[2] F. Acernese *et al.* (Virgo Collaboration), Advanced Virgo: A second-generation interferometric gravitational wave detector, Classical Quantum Gravity **32**, 024001 (2015).
[3] T. Akutsu *et al.* (KAGRA Collaboration), Overview of KAGRA: Detector design and construction history, Prog. Theor. Exp. Phys. **2021**, 05A101 (2021).
[4] B. P. Abbott *et al.* (Virgo and LIGO Scientific Collaborations), Observation of gravitational waves from a binary black hole merger, Phys. Rev. Lett. **116**, 061102 (2016).
[5] R. Abbott *et al.* (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, Phys. Rev. X **13**, 041039 (2023).
[6] A. H. Nitz, S. Kumar, Y.-F. Wang, S. Kastha, S. Wu, M. Schäfer, R. Dhurkunde, and C. D. Capano, 4-OGC: Catalog of gravitational waves from compact binary mergers, Astrophys. J. **946**, 59 (2023).
[7] S. Olsen, T. Venumadhav, J. Mushkin, J. Roulet, B. Zackay, and M. Zaldarriaga, New binary black hole mergers in the LIGO-Virgo O3a data, Phys. Rev. D **106**, 043009 (2022).
[8] E. D. Feigelson and G. J. Babu, Statistical methods for astronomy, arXiv:1205.2064.
[9] R. Abbott *et al.*, Population of merging compact binaries inferred using gravitational waves through GWTC-3, Phys. Rev. X **13**, 011048 (2023).
[10] C. Messick *et al.*, Analysis framework for the prompt discovery of compact binary mergers in gravitational-wave data, Phys. Rev. D **95**, 042001 (2017).
[11] S. 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.
[12] L. Tsukada *et al.*, Improved ranking statistics of the GstLAL inspiral search for compact binary coalescences, Phys. Rev. D **108**, 043004 (2023).
[13] K. Cannon *et al.*, GstLAL: A software framework for gravitational wave discovery, SoftwareX **14**, 100680 (2021).
[14] B. Ewing *et al.*, Performance of the low-latency GstLAL inspiral search towards LIGO, Virgo, and KAGRA's fourth observing run, Phys. Rev. D **109**, 042008 (2024).
[15] S. Sakon *et al.*, Template bank for compact binary mergers in the fourth observing run of Advanced LIGO, Advanced Virgo, and KAGRA, Phys. Rev. D **109**, 044066 (2024).
[16] 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).
[17] 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).
[18] B. Allen, χ^2 time-frequency discriminator for gravitational wave detection, Phys. Rev. D **71**, 062001 (2005).
[19] T. 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).
[20] S. A. Usman *et al.*, The PyCBCsearch for gravitational waves from compact binary coalescence, Classical Quantum Gravity **33**, 215004 (2016).
[21] A. H. Nitz, T. Dent, T. Dal Canton, S. Fairhurst, and D. A. Brown, Detecting binary compact-object mergers with gravitational waves: Understanding and improving the sensitivity of the PyCBC search, Astrophys. J. **849**, 118 (2017).
[22] G. S. Davies, T. Dent, M. Tápai, I. Harry, C. McIsaac, and A. H. Nitz, Extending the PyCBC search for gravitational waves from compact binary mergers to a global network, Phys. Rev. D **102**, 022004 (2020).
[23] J. Luan, S. Hooper, L. Wen, and Y. Chen, Towards low-latency real-time detection of gravitational waves from compact binary coalescences in the era of advanced detectors, Phys. Rev. D **85**, 102002 (2012).
[24] Q. Chu *et al.*, SPIIR online coherent pipeline to search for gravitational waves from compact binary coalescences, Phys. Rev. D **105**, 024023 (2022).
[25] 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).
[26] S. M. Gaebel, J. Veitch, T. Dent, and W. M. Farr, Digging the population of compact binary mergers out of the noise, Mon. Not. R. Astron. Soc. **484**, 4008 (2019).

[27] S. Galaudage, C. Talbot, and E. Thrane, Gravitational-wave inference in the catalog era: Evolving priors and marginal events, Phys. Rev. D **102**, 083026 (2020).
[28] J. Roulet, T. Venumadhav, B. Zackay, L. Dai, and M. Zaldarriaga, Binary black hole mergers from LIGO/Virgo O1 and O2: Population inference combining confident and marginal events, Phys. Rev. D **102**, 123022 (2020).
[29] J. Heinzel, C. Talbot, G. Ashton, and S. Vitale, Inferring the astrophysical population of gravitational wave sources in the presence of noise transients, Mon. Not. R. Astron. Soc. **523**, 5972 (2023).
[30] S. Banagiri, C. P. L. Berry, G. S. Cabourn Davies, L. Tsukada, and Z. Doctor, Unified p_{astro} for gravitational waves: Consistently combining information from multiple search pipelines, Phys. Rev. D **108**, 083043 (2023).
[31] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World* (Springer, New York, 2005), Vol. 29.
[32] L. S. Finn, Detection, measurement and gravitational radiation, Phys. Rev. D **46**, 5236 (1992).
[33] S. S. Chaudhary *et al.*, Low-latency gravitational wave alert products and their performance in anticipation of the fourth LIGO-Virgo-KAGRA observing run, Proc. Natl. Acad. Sci. U.S.A. **121**, e2316474121 (2024).
[34] D. Davis and M. Walker, Detector characterization and mitigation of noise in ground-based gravitational-wave interferometers, Galaxies **10**, 12 (2022).
[35] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), A guide to LIGO-Virgo detector noise and extraction of transient gravitational-wave signals, Classical Quantum Gravity **37**, 055002 (2020).
[36] B. S. Sathyaprakash and S. V. Dhurandhar, Choice of filters for the detection of gravitational waves from coalescing binaries, Phys. Rev. D **44**, 3819 (1991).
[37] B. Abbott *et al.* (LIGO Scientific Collaboration), Search for gravitational waves from binary inspirals in S3 and S4 LIGO data, Phys. Rev. D **77**, 062002 (2008).
[38] M. Was, M.-A. Bizouard, V. Brisson, F. Cavalier, M. Davier, P. Hello, N. Leroy, F. Robinet, and M. Vavoulidis, On the background estimation by time slides in a network of gravitational wave detectors, Classical Quantum Gravity **27**, 015005 (2010).
[39] 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 (2015).
[40] S. 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).
[41] T. Dent, Extending the PyCBC pastro calculation to a global network, IGFAE, University of Santiago de Compostela, Technical Report No. DCC-T2100060, LIGO, 2021.
[42] N. 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).
[43] A. Ray *et al.*, When to point your telescopes: Gravitational wave trigger classification for real-time multi-messenger followup observations, arXiv:2306.07190.
[44] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), 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).
[45] A. N. Angelopoulos and S. Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification, arXiv:2107.07511.
[46] G. Shafer and V. Vovk, A tutorial on conformal prediction, J. Mach. Learn. Res. **9**, 371 (2008).
[47] G. Ashton, Data release for: "Calibrating gravitational-wave search algorithms with conformal prediction" (Zenodo, 2023), 10.5281/zenodo.10246396.
[48] G. Cowan, *Statistical Data Analysis* (Oxford University Press, New York, 1998).
[49] V. Vovk, Conditional validity of inductive conformal predictors, Mach. Learn. **92** (2013).
[50] T. Ding, A. N. Angelopoulos, S. Bates, M. I. Jordan, and R. J. Tibshirani, Class-conditional conformal prediction with many classes, CoRR **abs/2306.09335** (2023).
[51] N. Colombo, On training locally adaptive *CP*, in *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, Proceedings of Machine Learning Research Vol. 204, edited by H. Papadopoulos, K. A. Nguyen, H. Boström, and L. Carlsson (PMLR, 2023), pp. 384–398.
[52] C. McIsaac and I. Harry, Using machine learning to autotune chi-squared tests for gravitational wave searches, Phys. Rev. D **105**, 104056 (2022).
[53] T. Venumadhav, B. Zackay, J. Roulet, L. Dai, and M. Zaldarriaga, New binary black hole mergers in the second observing run of Advanced LIGO and Advanced Virgo, Phys. Rev. D **101**, 083030 (2020).
[54] R. Abbott *et al.* (LIGO Scientific, KAGRA, and VIRGO Collaborations), Observation of gravitational waves from two neutron star–black hole coalescences, Astrophys. J. Lett. **915**, L5 (2021).