Enhancing Gravitational-Wave Detection: A Machine Learning Pipeline Combination Approach with Robust Uncertainty Quantification

Gregory Ashton,[1,*] Ann-Kristin Malz,[1] and Nicolo Colombo[2]
[1]*Department of Physics, Royal Holloway, University of London, Egham Hill, Egham TW20 0EX, United Kingdom*
[2]*Department of Computer Science, Royal Holloway University of London, Egham Hill, Egham TW20 0EX, United Kingdom*

(Received 25 April 2025; accepted 4 December 2025; published 8 January 2026)

Gravitational-wave data from advanced-era interferometric detectors consists of background Gaussian noise, frequent transient artifacts, and rare astrophysical signals. Multiple search algorithms exist to detect the signals from compact binary coalescences, but their varying performance complicates interpretation. We present a machine-learning-driven approach that combines results from individual pipelines and utilizes conformal prediction to provide robust, calibrated uncertainty quantification. Using simulations, we demonstrate improved detection efficiency and apply our model to GWTC-3, enhancing confidence in multipipeline detections, such as the subthreshold binary neutron star candidate GW200311_103121.

DOI: 10.1103/yfb3-fgf2

Gravitational-wave astronomy is progressing from initial detection to routine observation. As of GWTC-4.0, the fourth gravitational-wave transient catalog [1], the LIGO Scientific, Virgo, and KAGRA (LVK) Collaborations have detected over 200 signals arising from compact binary coalescence (CBC) sources. These sources are discovered using highly developed search algorithms (pipelines) that measure significance by comparing a detection statistic for the candidate against an empirical background distribution. Candidates are initially ranked by a frequentist false alarm rate (FAR), but the pipeline outputs are then convolved with an astrophysical model of the CBC population to produce a Bayesian $p_{\rm astro}$ [2–5]. The LVK routinely uses five pipelines to detect signals. Four of these (GstLAL [6–11], MBTA [12,13], PyCBC [14–18], and SPIIR [19,20]) use parametrized models of CBC sources, while cWB [21] uses a wavelet model with weaker assumptions about the source type. Moreover, there are also external teams that run independent searches (see, e.g., Refs. [22,23]).

To date, a straightforward approach has been taken to combining the results from multiple pipelines: taking the maximum $p_{\rm astro}$ or inverse FAR (IFAR: i.e., 1/FAR) across the set of contributing pipelines. For example, in the GWTC, candidates with at least one pipeline with $p_{\rm astro} > 0.5$ are considered significant signals (with an estimated contamination rate from nonastrophysical sources of 10%–15%, see, e.g., Abbott *et al.* [24]). This powerful and straightforward approach does not require processing and enables the simple combination of independent catalogs. However, multiple estimates of a candidate's significance and properties by different algorithms also present an opportunity: correlations between pipelines could be exploited to improve the overall detection efficiency beyond the current maximum approach. This idea has already been explored for combining the pipelines to produce a unified $p_{\rm astro}$ [25], but this relies on accurate models of the signal and noise distributions. In this Letter, we explore a new approach that combines pipelines using simple machine learning (ML) models trained on labeled data. However, the predictions from such models are uncalibrated and lack a quantified uncertainty. Therefore, we augment our ML-combination pipeline by applying conformal prediction (CP) [26,27] to provide quantified uncertainty measurements using labeled calibration data. This approach is fast, computationally efficient, and requires only the simulation of the expected signal and noise distributions. Our approach offers the capacity to learn the strengths and weaknesses of multiple pipelines without strict requirements on the underlying data products. Thus, it can also be used to assess the performance of new pipelines or modifications to existing pipelines. We restrict ourselves to the binary classification problem, signal or noise, but the work can be generalized to multiclass classification straightforwardly.

We use two standard ML classification models: logistic regression (LR) and a multilayer perceptron (MLP); both are discussed in detail in Supplemental Material [28]. Each takes as input a feature vector $\vec{X}$ and returns a normalized set of probabilities P for each label. To train the models, we utilize the results from the recent mock data challenge (MDC) study in advance of the LVK fourth observing run [29], where the GstLAL, PyCBC, SPIIR, MBTA, and cWB search

[*]Contact author: gregory.ashton@ligo.org

pipelines were applied to a real-time data replay with added simulated signals. The 40 days of data are taken from the LIGO Livingston and Hanford detectors [30] and the Virgo detector [31] during the third observing run (the KAGRA detector [32] was not in operation at this time). From this MDC, we take all candidates, excluding early warning candidates, that the search pipelines upload, cluster in time (grouping all events within a 1 s window), and then filter all but the maximum signal-to-noise ratio (SNR) candidate per pipeline.

This produces our feature data $\{\vec{X}_n\}$ where each row contains the per-pipeline features (including detection quantities such as the IFAR, SNR alongside estimates of the source properties such as the mass and spin; see Supplemental Material [28] for details). However, we do not include $p_{\rm astro}$ as a feature because the enhanced signal rate used in constructing the data mean the pipeline $p_{\rm astro}$ values are not well calibrated. For elements of the feature data where any given pipeline does not find a candidate with IFAR > 1 h, we enter zeros to fill in the missing data.

We then compare the candidate list with the times of known simulated signals and astrophysical signals known to be in the data and produce a ground-truth label set $\{Y_n\}$. We find 9946 rows of data, with 5908 corresponding to simulated or real signals. The number of signals in this data is significantly greater than the rate of detections expected for advanced-era detectors, as simulated signals were added to the data at a rate much higher than the anticipated astrophysical rate to stress test the low-latency infrastructure. With the data in hand, we then split the data into three subsets: 10% for CP calibration, 10% for testing, and the remainder for training.

In Fig. 1, we compare the receiver operator curve (ROC) for the two ML pipeline combination approaches to the standard maximum-IFAR method. This demonstrates the potential of ML to improve the detection efficiency, as quantified by the area under curve (AUC) provided in the legend. Specifically, both ML approaches deliver an increase in the AUC above the level of uncertainties in the AUC as measured over the test data. However, we note that comparing the uncertainty in the ROC curve itself, the distributions do overlap, but their uncertainty envelopes are visually separated.

Comparing the two ML approaches, Fig. 1 demonstrates that the MLP approach outperforms the simpler LR model as measured by the AUC. This is expected since the MLP is more expressive: it can capture more complicated patterns due to the more involved underlying architecture. However, while the LR model is simpler, the results are easily interpreted. A simple inspection of the fitted coefficients can provide insight into the importance of individual features for each pipeline (see Supplemental Material [28]). For the advantage of interpretability and a modest reduction in performance, we present only the results for the LR model hereafter.

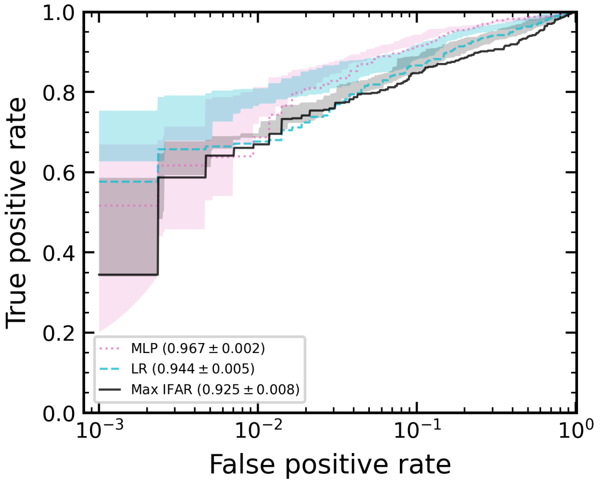

FIG. 1. The ROC for the LR and MLP ML-driven pipeline combination approaches applied to the test data; we also include as a comparison the standard maximum-IFAR pipeline combination approach. For this test, we use all four pipelines contributing to the MDC and all features in our test data (see Supplemental Material [28] for details). To investigate the uncertainty inherent in the ROC curve, we run the study under different permutations of the training and test data. The solid lines indicate the ROC calculated for a single permutation of the test data, while the shaded band marks the 90% interval over the permutations. We quantify the difference between combination approaches in the legend by providing the AUC along with an estimate calculated under several training and test data permutations.

So far, we have demonstrated that an ML-driven pipeline combination approach can outperform a naive maximum-IFAR combination as measured by the ROC. However, in contrast to the results from combining individual search pipeline results, an ML pipeline combination does not provide a well-calibrated measure of the uncertainty. This is important because a central aim for any method seeking to identify signals is to assess the significance of individual candidates. As argued in Gebhard *et al.* [33], discussed in the context of using a convolutional neural network (CNN) as a search algorithm, the output of any ML classifier is a function of the test data set and therefore is not necessarily calibrated to reality. They conclude that "CNNs alone cannot be used to properly claim gravitational-wave detections."

This difficulty is not unique to gravitational-wave astronomy; uncertainty quantification is a topic of interest in many high-stakes applications of ML where predictions must be robust. One approach to providing robust, well-calibrated predictions is CP, a distribution-free approach that requires only exchangeability of the data and can be applied to any point predictor to produce statistically rigorous prediction regions [26,27]. In Supplemental Material [28], we provide a brief introduction to CP, but we have previously applied CP to the problem of gravitational-wave astronomy [34], demonstrating how to calibrate individual pipelines. We now extend that work to quantify uncertainty for an ML combination pipeline. Specifically, we apply standard label-conditional prediction using the complement of the LR prediction probability as a nonconformity score (in Malz *et al.* [35], we explore alternative scores but do not find compelling advantages to use these in this Letter).

To measure the significance of an individual event within the CP framework, we can use the confidence [36]. In Ashton *et al.* [34], we explored three possible definitions of the confidence, each with its own merits. In this Letter, we will apply the "conditional confidence: signal" which is defined as the minimum value of α such that the signal label is included in the prediction set Γ^{α}. We choose this conditional confidence because (i) unlike the standard definition of the confidence [36], it can be measured for any label on any test data, (ii) it can be generalized to the multiclass case trivially, and (iii) it enables the straightforward definition of a catalog with a calculable purity by placing a threshold on the conditional confidence.

In Fig. 2, we plot the conditional confidence against the maximum IFAR for all test data points using the LR model and then highlight the true label, number of contributing pipelines, and measured chirp mass [37] of the signal. Comparing the confidence with the maximum IFAR, we observe that events identified by multiple pipelines tend to have higher confidence (events found by a single pipeline are mostly located at the lower edge of the distribution for a given IFAR). Examining specific events, we observe a single false positive (using a standard threshold of 1 per year), which is assigned an IFAR of approximately 10^{10} s by the GstLAL pipeline. Under the maximum-IFAR approach, such an event would be considered significant; however, the confidence of the event is found to be small ($\approx$0.35) relative to other candidates found by multiple pipelines at a similar IFAR. However, on the other hand, nearby this candidate, there is a low-mass event found by GstLAL and PyCBC, which is ranked with a similar confidence despite being found by multiple pipelines. This suggests more work is needed to understand how the confidence is assigned and optimize it to better separate signals and noise.

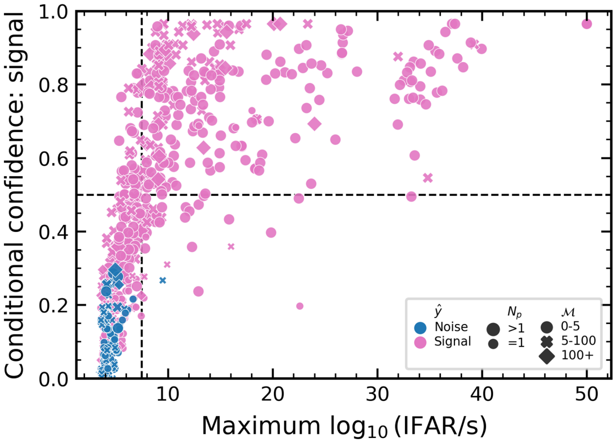

FIG. 2. The conditional confidence in the signal label as measured by the LR model and applied to the test data compared to the maximum IFAR. We highlight the true label ($\hat{y}$) by the color, the number of contributing pipelines (N_p) by the size, and the chirp mass ($\mathcal{M}$) range inferred by the highest-SNR pipeline by the symbol. A vertical dashed line marks a FAR threshold of 1 per year.

A core assumption of CP is that the calibration data and the test data are *exchangeable* [36]: given a collection of N data points, the N different orderings are equally likely. For the problem of reasonably well-calibrated search pipelines studying a stream of data (e.g., from a given observing run), it seems reasonable that their results will be exchangeable. I.e., we do not expect the meaning of the FAR and the measured parameters, such as mass, to vary throughout an observing run. However, this may not be true if there are changes to the search pipeline or the instruments (say, we utilize examples from a previous observing run). Therefore, care should be taken whenever the calibration data and test data are sourced differently (which will always be the case when studying real data, as we do not know the ground truth about the sources impacting our detectors). Moreover, careful investigation is needed to understand the importance of the relative numbers of signals and noise candidates in the calibration data. For the demonstrations above, we have guaranteed exchangeability by randomly splitting the MDC data set. We will now go beyond this test data set to study results from real searches for signals.

We study the candidate lists from the O3a and O3b observing runs published as part of the GWTC-3 catalog [38,39]. Specifically, this includes a list of the search pipeline output from the CWB, PyCBC, GstLAL, and MBTA pipelines. However, for PyCBC, a second search was performed targeting only binary black hole (BBH) candidates; we excluded these results to improve the exchangeability of the training and test data. Furthermore, we utilize only the measured IFAR, SNR, and chirp mass (except for CWB); this is done to best ensure exchangeability, as the FAR is calibrated and checked during pipeline development. Nevertheless, we acknowledge that the pipelines were developed between O3 analyses and the MDC, so there are likely differences in their behavior. We take MDC data using this restricted feature set for use in training the LR model and for calibration.

In Fig. 3, we plot the conditional confidence obtained from our LR combination model against the maximum p_{astro} across pipelines. We compare against p_{astro} here as this is the primary metric used in the GWTC to threshold for further analysis. However, we note that p_{astro} is not included in the features used to train our LR model. This is because, in addition to the issues with the p_{astro} values within the MDC [29], while it is possible to use p_{astro} as a

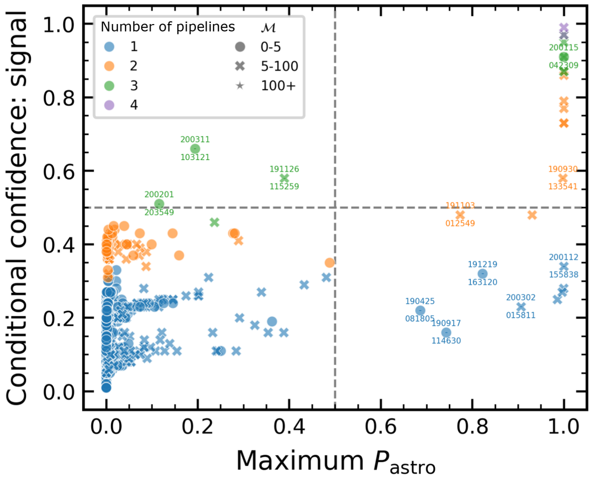

FIG. 3. A comparison of the p_{astro} and conditional confidence using the LR model trained on a subset of the MDC data.

feature, this is one of the features we know can be nonexchangeable since the astrophysical population improves as we see more events. Therefore, the population model used to calculate p_{astro} for the training data is different from that used to calculate p_{astro} for the test data. As a result, the p_{astro} presented in Fig. 3, contains information about the astrophysical population not available in the measurement of the conditional confidence.

From Fig. 3, the four quadrants reveal an insight into the comparative performance of the traditional p_{astro} method and the CP confidence. First, we note that they are correlated: we have most of the data points in the top right and bottom left. In the top-right quadrant, we see a cluster of events with a $p_{\text{astro}} \sim 1$ and confidence ∼1 (see Supplemental Material [28]). Just below this cluster. we also find GW200115_042309, one of the first detected neutron star black hole (NSBH) signals [40] with a confidence of ∼0.9, which was not detected by cWB. Finally, we also find GW200209_085452, a BBH candidate not found by cWB. In the bottom-left quadrant, we find subthreshold candidates from both methods; we observe some stratification in the confidence, which is not yet understood.

In the bottom-right quadrant, we find candidates with $p_{\text{astro}} > 0.5$ but confidence below 0.5. Except in two cases with a confidence level of ∼0.5, these candidates are identified by only a single pipeline. For example, GW200302_015811 was found by GstLAL in data from Hanford and Virgo, but the Virgo data had an SNR less than 4. Meanwhile, GW200220_061928 is a high-mass candidate found only by PyCBC. The candidate with the lowest confidence but highest p_{astro} is GW190917_114630, found only by GstLAL in GWTC-2.1 with a p_{astro} of ∼0.7 [24,41]. Based on the source properties, this is most likely an NSBH [41,42]. However, its properties are also found to be inconsistent with the isolated binary evolution pathway [43]. Nearby this event, we also find GW190425_081805, the second observed binary neutron star (BNS) [44], which is similarly only found by the GstLAL pipeline (again, we report the updated p_{astro} from GWTC-2.1 [41]).

Finally, we focus on the upper-left-hand quadrant: candidates above a confidence threshold of 0.5 but below a p_{astro} of 0.5, where we find three candidates. First, GW191126_115259 is a BBH candidate found by GstLAL, PyCBC, and MBTA with a maximum p_{astro} of 0.39 (in PyCBC), but a confidence of ∼0.6. Notably, this event is detected by the PyCBC -BBH search with a p_{astro} of 0.7 (these results are excluded from our test data set for reasons stated above). Next, GW200311_103121 and GW200201_203549 both appear in the marginal candidate table of GWTC-3, and from a multicomponent p_{astro} analysis, they are indicated (if real) to be a BNS and NSBH, respectively. These events are given greater confidence relative to their maximum p_{astro}, which arises from the fact that three pipelines find them. While we do not claim that our new metric is robustly tested enough to claim these as up-ranked detections, they demonstrate the increased significance possible from combining pipelines. A reanalysis of this data, with better control over any systematic differences in pipeline behavior, may yield the ability to trust the increased significance. If confirmed, GW200311_103121 would be only the third BNS signal detected, underscoring the importance of using all available information to assess significance.

In summary, we have introduced a new approach to pipeline combination, offering improvements over the current method that takes a simple maximum p_{astro} or IFAR. In this approach, we utilize ML to learn the optimal combination from a set of training data in which the ground truth is known. Utilizing a recent MDC [29], we demonstrated that simple off-the-shelf LR or MLP models can outperform the maximum-IFAR combination at the population level. However, such an approach is limited by itself, as it lacks a robust measurement of prediction uncertainty for individual events. Therefore, we introduce CP, which can provide robust uncertainty measurements through an additional calibration data set. We demonstrate the application of the CP confidence to observed data from O3 and find a handful of events (most notably GW200311_103121, a possible BNS event), where the combination approach yields increased confidence relative to p_{astro}, which stems from the multiple pipelines that identified the candidate.

For experts within the field, utilizing the outputs from multiple pipelines is a standard process when assessing the significance of a candidate. The combination approach proposed here does not replace that expertise but aims to enhance it, providing a single rigorous quantified uncertainty. The key beneficiary of this approach is astronomers who use gravitational-wave alerts and the GWTC (e.g., to trigger observations or perform further studies). A single statement of confidence in a candidate, which combines the parameter-space-dependent sensitivity of all pipelines, will provide clarity and interpretability. We envision that the field could utilize this approach to combine search pipelines when producing a catalog of events or low-latency alerts. A single, easy-to-understand assessment of the multipipeline results will ease interpretation and potentially improve sensitivity.

Implementation of our ML models was done using SCIKIT-LEARN [45], while we utilize NumPy [46], Pandas [47], and Matplotlib [48] for data handling and visualization.

Acknowledgments—We thank Sharan Banagiri, Francesco Salemi, Tito Dal Canton, and Surabhi Sachdev for discussions that helped in the development of this work and the two anonymous referees who provided useful feedback during the review. We also acknowledge unpublished work by Nikolas Moustakidis, Theofilos Moustakidis, Erik Katsavounidis, and Deep Chatterjee that took a similar approach to our ML pipeline combination method. 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. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants No. PHY-0757058 and No. PHY-0823459. This work is supported by the Science and Technology Facilities Council (STFC) Grant No. UKRI2488.

Data availability—The data that support the findings of this article are not publicly available because they are owned by a third party and the terms of use prevent public distribution. The data are available from the authors upon reasonable request.

[1] A. G. Abac *et al.*, GWTC-4.0: Updating the Gravitational-Wave Transient Catalog with Observations from the First Part of the Fourth LIGO-Virgo-KAGRA Observing Run, Report No. LIGO-P2400386, 2025, https://inspirehep.net/literature/2963698.

[2] 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).

[3] T. Dent, Extending the PyCBC pastro calculation to a global network, Technical Report No. DCC-T2100060, LIGO, 2021, https://dcc.ligo.org/LIGO-T2100060/public.

[4] 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).

[5] Anarya Ray *et al.*, When to Point Your Telescopes: Gravitational Wave Trigger Classification for Real-Time Multi-Messenger Followup Observations, Report No. LIGO-P2300141, 2023, https://inspirehep.net/literature/2668083.

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

[7] 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.

[8] Leo Tsukada *et al.*, Improved ranking statistics of the GstLAL inspiral search for compact binary coalescences, Phys. Rev. D **108**, 043004 (2023).

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

[10] Becca 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).

[11] Shio 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).

[12] 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).

[13] 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).

[14] Bruce Allen, χ^2 time-frequency discriminator for gravitational wave detection, Phys. Rev. D **71**, 062001 (2005).

[15] 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).

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

[17] 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).

[18] 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).

[19] Jing Luan, Shaun Hooper, Linqing Wen, and Yanbei 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).

[20] Qi Chu *et al.*, SPIIR online coherent pipeline to search for gravitational waves from compact binary coalescences, Phys. Rev. D **105**, 024023 (2022).

[21] S. Klimenko *et al.*, Method for detection and reconstruction of gravitational wave transients with networks of advanced detectors, Phys. Rev. D **93**, 042004 (2016).

[22] 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, Astrophys. J. **946**, 59 (2023).

[23] Ajit Kumar Mehta, Seth Olsen, Digvijay Wadekar, Javier Roulet, Tejaswi Venumadhav, Jonathan Mushkin, Barak Zackay, and Matias Zaldarriaga, New binary black hole mergers in the LIGO-Virgo O3b data, Phys. Rev. D **111**, 024049 (2025).

[24] R. Abbott *et al.*, 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).

[25] Sharan Banagiri, Christopher P. L. Berry, Gareth S. Cabourn Davies, Leo Tsukada, and Zoheyr Doctor, Unified pastro for gravitational waves: Consistently combining information from multiple search pipelines, Phys. Rev. D **108**, 083043 (2023).

[26] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer, *Algorithmic Learning in a Random World* (Springer, New York, 2005), Vol. 29.

[27] Anastasios N. Angelopoulos and Stephen Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification, arXiv:2107.07511.

[28] See Supplemental Material at http://link.aps.org/supplemental/10.1103/yfb3-fgf2 for additional studies and details of the algorithm.

[29] Sushant Sharma Chaudhary *et al.*, Low-latency gravitational wave alert products and their performance in anticipation of the fourth LIGO-Virgo-KAGRA observing run, Proc. Nat. Acad. Sci. **121**, e2316474121 (2025).

[30] J. Aasi *et al.*, Advanced LIGO, Classical Quantum Gravity **32**, 074001 (2015).

[31] F. Acernese *et al.*, Advanced Virgo: A second-generation interferometric gravitational wave detector, Classical Quantum Gravity **32**, 024001 (2015).

[32] T. Akutsu *et al.*, KAGRA: 2.5 generation interferometric gravitational wave detector, Nat. Astron. **3**, 35 (2019).

[33] Timothy D. Gebhard, Niki Kilbertus, Ian Harry, and Bernhard Schölkopf, Convolutional neural networks: A magic bullet for gravitational-wave detection?, Phys. Rev. D **100**, 063015 (2019).

[34] Gregory Ashton, Nicolo Colombo, Ian Harry, and Surabhi Sachdev, Calibrating gravitational-wave search algorithms with conformal prediction, Phys. Rev. D **109**, 123027 (2024).

[35] Ann-Kristin Malz, Gregory Ashton, and Nicolo Colombo, Classification uncertainty for transient gravitational-wave noise artefacts with optimised conformal prediction, Phys. Rev. D **111**, 084078 (2025).

[36] Glenn Shafer and Vladimir Vovk, A tutorial on conformal prediction, J. Mach. Learn. Res. **9**, 371 (2008).

[37] A. G. Abac *et al.*, GWTC-4.0: An introduction to version 4.0 of the gravitational-wave transient catalog, Astrophys. J. Lett. **995**, L18 (2025).

[38] LIGO Scientific Collaboration and Virgo Collaboration, Gwtc-2.1: Deep extended catalog of compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run—candidate data release, 10.5281/zenodo.5759108 (2021).

[39] 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—candidate data release, 10.5281/zenodo.5546665 (2021).

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

[41] R. Abbott *et al.*, GWTC-2.1: Deep extended catalog of compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run, Phys. Rev. D **109**, 022001 (2025).

[42] R. Abbott *et al.*, Population of merging compact binaries inferred using gravitational waves through GWTC-3, Phys. Rev. X **13**, 011048 (2023).

[43] Floor S. Broekgaarden and Edo Berger, Formation of the first two black hole–neutron star mergers (GW200115 and GW200105) from isolated binary evolution, Astrophys. J. Lett. **920**, L13 (2021).

[44] B. P. Abbott *et al.*, GW190425: Observation of a compact binary coalescence with total mass $\sim 3.4 M_{\odot}$, Astrophys. J. Lett. **892**, L3 (2020).

[45] Lars Buitinck *et al.*, API design for machine learning software: Experiences from the SCIKIT-LEARN project, arXiv:1309.0238.

[46] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith *et al.*, Array programming with NumPy, Nature (London) **585**, 357 (2020).

[47] The Pandas Development Team, Pandas-dev/Pandas: Pandas, 10.5281/zenodo.3509134 (2020).

[48] John D. Hunter, Matplotlib: A 2d graphics environment, Comput. Sci. Eng. **9**, 90 (2007).