Deep learning and Bayesian inference of gravitational-wave populations: Hierarchical black-hole mergers

Matthew Mould,[1,*] Davide Gerosa,[2,3,1] and Stephen R. Taylor[4]

[1]*School of Physics and Astronomy and Institute for Gravitational Wave Astronomy,*
University of Birmingham, Birmingham B15 2TT, United Kingdom
[2]*Dipartimento di Fisica "G. Occhialini", Universitá degli Studi di Milano-Bicocca,*
Piazza della Scienza 3, 20126 Milano, Italy
[3]*INFN, Sezione di Milano-Bicocca, Piazza della Scienza 3, 20126 Milano, Italy*
[4]*Department of Physics and Astronomy, Vanderbilt University,*
2301 Vanderbilt Place, Nashville, Tennessee 37235, USA

(Received 8 March 2022; accepted 7 October 2022; published 10 November 2022)

The catalog of gravitational-wave events is growing, and so are our hopes for constraining the underlying astrophysics of stellar-mass black-hole mergers by inferring the distributions of, e.g., masses and spins. While conventional analyses parametrize this population with simple phenomenological models, we propose an emulation-based approach that can compare astrophysical simulations against gravitational-wave data. We combine state-of-the-art deep-learning techniques with hierarchical Bayesian inference and exploit our approach to constrain the properties of repeated black-hole mergers from the gravitational-wave events in the most recent LIGO/Virgo catalog. Deep neural networks allow us to (i) construct a flexible single-channel population model that accurately emulates simple parametrized numerical simulations of hierarchical mergers, (ii) estimate selection effects, and (iii) recover the branching ratios of repeated-merger generations. Among our results, we find the following: The distribution of host-environment escape speeds favors values less than 100 $\mathrm{km\,s^{-1}}$ but is relatively flat, with around 37% of first-generation mergers retained in their host environments; first-generation black holes are born with a maximum mass that is compatible with current estimates from pair-instability supernovae; there is multimodal substructure in both the mass and spin distributions, which, in our model, can be explained by repeated mergers; and binaries with a higher-generation component make up at least 14% of the underlying population. Though these results are inferred through emulation of a simplified model, the deep-learning pipeline we present is readily applicable to realistic astrophysical simulations.

DOI: 10.1103/PhysRevD.106.103013

I. INTRODUCTION

The Advanced LIGO [1] and Virgo [2] gravitational-wave (GW) detectors are revealing the previously unseen landscape of compact binary coalescences. To date, nearly 100 GW signals from merging stellar-mass compact objects have been observed, the majority being black holes (BHs) [3–9]. Accurate estimation of the intrinsic properties of individual sources, such as component masses and spins, allows us to view the distribution of merging binary BHs as a whole. Crucially, the binary parameters inferred at merger are influenced by the formation history and astrophysical environment in which the progenitor systems were born; conversely, cumulative measurements of those source properties allow constraints to be placed at the population level, which can ultimately be compared to the predictions from likely binary formation scenarios.

Two examples include isolated stellar binary evolution [10] and dynamical interactions in star clusters [11]. While the former predicts a forbidden mass region for stellar remnants [12,13] and spins that favor small misalignments with the binary orbital angular momentum [14–16], binaries formed in the latter channel may repeatedly interact and merge with other members of the cluster and thus be pushed to higher masses and isotropic spin orientations [17] (with GW-driven inspiral preserving the spin isotropy [18,19]).

Given the catalog of GW detections, one can take two approaches to assess the underlying astrophysical population of binary BHs. In a simulation-based analysis, sources are synthesized—accounting for as many detailed astrophysical processes as are known or are computationally feasible—to form distributions of detectable merging binaries. By varying population-level input parameters controlling binary evolution (e.g., common envelope efficiency and strength of supernova kicks), one can assess the degree of consistency with the observed events (for reviews, see,

[*] mmould@star.sr.bham.ac.uk

e.g., Refs. [20–22]). However, such simulations are typically computationally intensive, and large uncertainties remain on key parameters (see, e.g., Refs. [23,24]).

The second approach is to first construct a model of the astrophysical distribution of source parameters—which is conditionally dependent on given population-level parameters controlling its shape (the "hyperparameters," e.g., mass cutoffs or spectral indices)—and use the observed catalog to perform a hierarchical Bayesian inference that accounts for observational biases (e.g., that heavier sources are easier to detect). This statistical analysis is hierarchical in the sense that one uses previous Bayesian measurements of the binary BH source parameters to then measure said hyperparameters [25,26]. The population model used could be as in the previous approach such that the distribution is known only at discrete values of the hyperparameters, but this would allow only for single posterior evaluations for relative comparisons (e.g., via Bayes factors) and leave some of the hyperparameters unconstrained (see, e.g., Refs. [27,28] for examples of this approach).

On the other hand, a population model that can be continuously evaluated across the population-level parameter space can be used to make Bayesian measurements of the hyperparameters. This requirement typically necessitates simple, quick-to-evaluate parametric forms with statistical independence between source parameters (see, e.g., the models used in Refs. [29–31]) to enable efficient hyperposterior sampling. The disadvantage of this approach is that it is inherently phenomenological with a discretionary selection of the employed functional forms. Recent work has sought to improve parametric population models by addressing potential correlations between mass and spin parameters [31–33] and assessing the suitability of spin parametrizations [34–36] since accurate inference requires appropriate models [37]. Along other lines, the flexibility of population analyses can be improved with semiparametric and nonparametric modeling techniques [38–42].

Previous studies have focused on combining the simulation-based and parametric approaches: A simulation emulator constructed with sufficient accuracy to rapidly synthesize predictive distributions over the hyperparameter space can be adopted in place of parametrized phenomenological models in the Bayesian inference pipeline. Such models leverage the advantages of efficient hyperposterior sampling and direct astrophysics-to-GW data comparison provided by each approach.

A first step in this direction within the context of GW population inference was taken by some of the authors in Ref. [43]. Compressed principal components of binned simulation data were emulated over low- (typically one- or two-) dimensional source- and population-level parameter spaces using Gaussian process regression (GPR). However, this emulation approach was shown to be unsuitable for extension to more complex higher-dimensional modeling scenarios due to poor predictive accuracy and infeasible computational requirements [44,45]. These issues were tackled in Ref. [46] by employing deep-learning techniques to construct simulation-informed population models; in particular, the conditional density estimator takes the form of a flow-based generative neural network known as a normalizing flow [47]. In general, neural networks are powerful tools that offer greater flexibility when employed as functional emulators. In this case, normalizing flows prompted population studies considering the scenarios of primordial BHs [48] and mixture models between isolated and dynamical evolution [49].

In this work, we develop complementary deep-learning techniques that build on the advancements of Refs. [43,46] by pushing the emulated parameter space dimensionality and introducing new neural network applications. We employ fully connected deep neural networks (DNNs; also referred to as multilayer perceptrons) to act as the conditional density estimator of a population model and to capture the effect of GW detection biases on the population of observed binary BH events (see also Refs. [50–52] for machine-learning approaches to estimating selection effects).

Motivated by evidence for large masses in the observed GW catalog, we apply these deep-learning techniques to binary BH populations containing hierarchical mergers, in which component BHs may be the remnants of (multiple) previous mergers [17]. These so-called "higher-generation" BHs may explain the outlier properties of events such as GW190412 [53–57], GW190521 [58–65] (though see also Refs. [66,67], which find that these events may in fact be consistent with the population), and GW190814 [68–71]. The presence of hierarchical mergers in binary BH populations is crucially dependent on the escape speeds of dynamical host environments (e.g., young star clusters, globular clusters, and nuclear star clusters [63]) and the magnitudes of gravitational recoils received due to the anisotropic emission of GWs [72–74].

Our DNN population model learns from simple simulations of clusterlike environments [75], which account for the retention and ejection of merger remnants due to GW kicks. We model the *joint* distribution of four source parameters—two masses and two effective spins, which present identifiable features due to the influence of higher-generation BHs [76,77]—and six population-level (hyper) parameters. These hyperparameters control the population properties of first-generation BHs born in stellar collapse, the binary pairing process, and host escape speeds. We also train a DNN to predict the fractional contributions of the population-dependent first-, mixed-, second-, and higher-generation binary BHs.

We illustrate our procedure schematically in Fig. 1, in which each element represents a single modeling process, arrows direct the one-way flow of information between them, and rows group distinct stages of our pipeline. The first row represents simulations, controlled

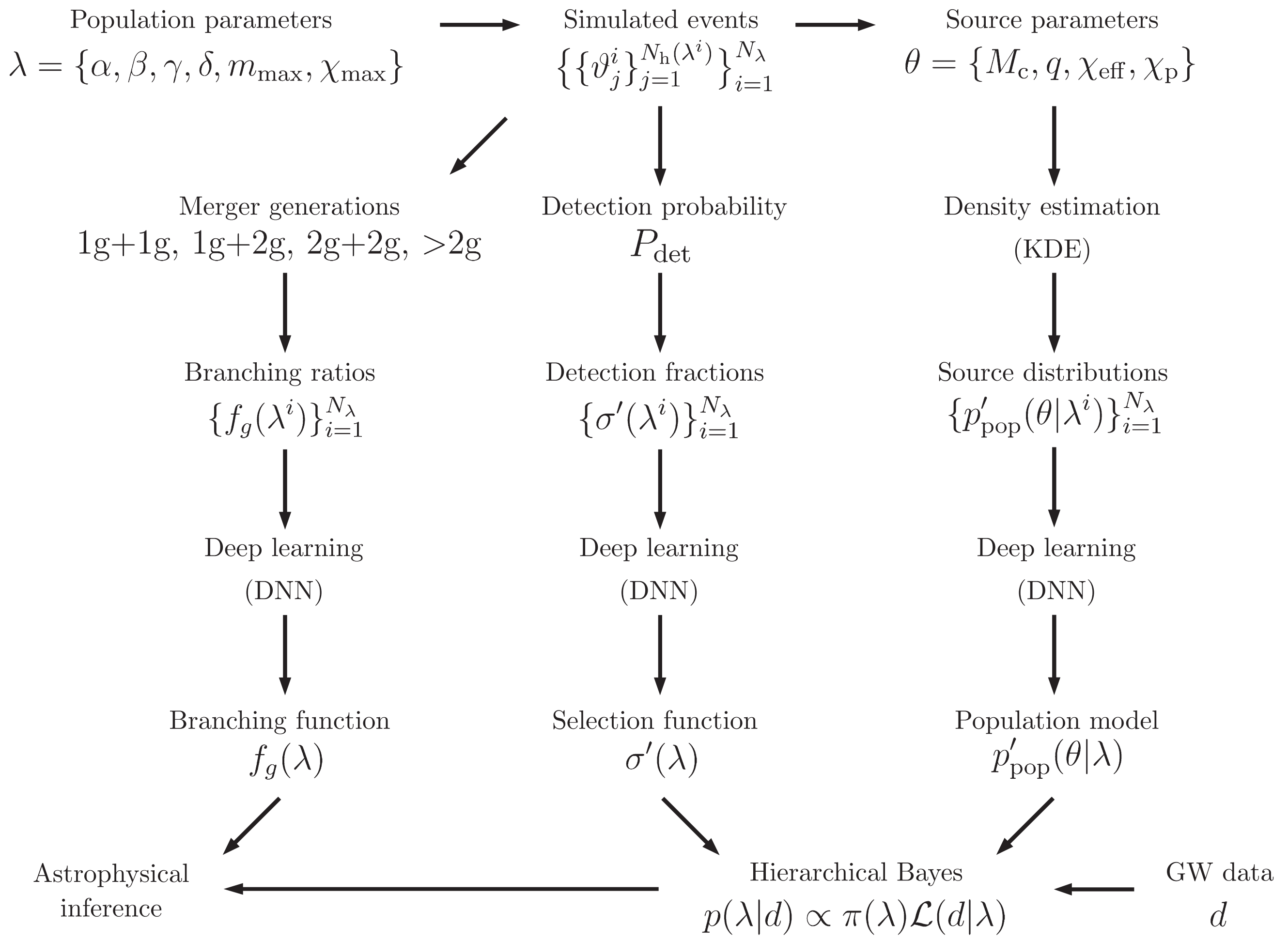
Population parameters
$\lambda = \{\alpha, \beta, \gamma, \delta, m_{\max}, \chi_{\max}\}$
Simulated events
$\{\{\vartheta_j^i\}_{j=1}^{N_{\mathrm{h}}(\lambda^i)}\}_{i=1}^{N_\lambda}$
Source parameters
$\theta = \{M_{\mathrm{c}}, q, \chi_{\mathrm{eff}}, \chi_{\mathrm{p}}\}$
Merger generations
1g+1g, 1g+2g, 2g+2g, >2g
Detection probability
P_{det}
Density estimation
(KDE)
Branching ratios
$\{f_g(\lambda^i)\}_{i=1}^{N_\lambda}$
Detection fractions
$\{\sigma'(\lambda^i)\}_{i=1}^{N_\lambda}$
Source distributions
$\{p'_{\mathrm{pop}}(\theta|\lambda^i)\}_{i=1}^{N_\lambda}$
Deep learning
(DNN)
Deep learning
(DNN)
Deep learning
(DNN)
Branching function
$f_g(\lambda)$
Selection function
$\sigma'(\lambda)$
Population model
$p'_{\mathrm{pop}}(\theta|\lambda)$
Astrophysical
inference
Hierarchical Bayes
$p(\lambda|d) \propto \pi(\lambda)\mathcal{L}(d|\lambda)$
GW data
d

FIG. 1. Schematic diagram of our population modeling and inference procedure. Arrows indicate information that is passed from one element to another, and elements that occur at the same stage of the pipeline are grouped into rows. The first row represents simulations of binary BH mergers, while the second lists postprocessing applied to the simulated data. We leverage deep learning, shown in the third row, by constructing DNNs to act as functional emulators for key ingredients of GW population inference, indicated within the fourth row. In the final row, the deep-learned selection function and population model are combined with data from GW catalogs to feed into a hierarchical Bayesian inference, which, along with a third DNN to predict branching fractions between subpopulations, is used to make conclusions about the underlying distribution of merging stellar-mass binary BHs.

by population-level parameters, of binary BH mergers characterized by a complete set of source-level parameters that are condensed into those we model. In the second row, we list the postprocessing performed on the simulated data. For each simulation, we construct the joint probability density of modeled source-level parameters conditioned on the population-level parameters, the expected fractional number of detectable sources, and the relative contributions from each hierarchical merger generation to the total population. We transform these discrete sets of evaluations into continuous functions using deep learning, as seen in the third row. These DNN functional emulators, listed in the fourth row, are employed in conjunction with data from the GW events detected to date to perform a hierarchical Bayesian inference and ultimately constrain the population of merging stellar-mass binary BHs, as illustrated in the final row. Each ingredient and the relevant symbols are defined throughout the paper.

In Sec. II, we describe our simple approach to generating sets of simulated hierarchical merger distributions. We lay out the statistical tools of population inference (Sec. III A), as well as our aforementioned use of DNNs to estimate population models (Sec. III B), selection biases (Sec. III C), and population-dependent branching fractions (Sec. III D). Our deep-learning-enhanced statistical pipeline is validated with mock GW catalogs in Sec. IV. In Sec. V, we report the results of our inference on the latest catalog of GW events, discussing the astrophysical implications and comparing to recent related works. We finish with a summary of future extensions to our work in Sec. VI and concluding remarks in Sec. VII. The GW events that are included in our analysis and their source parameters are enumerated in the Appendix.

The inference pipeline established here highlights advancements at the intersection of GW astronomy with statistical analysis and deep learning, and readily accommodates more realistic astrophysical simulations such as binary population synthesis.

II. HIERARCHICAL MERGER POPULATIONS

We model the retention and ejection of merger remnants in a "cluster," which here simply refers to a collection of BHs in an environment with constant escape speed $v_{\rm esc}$. We use the setup described in Ref. [75] (see Refs. [54,77] for additional applications). Our model depends on six population parameters, $\lambda \coloneqq \{\alpha, \beta, \gamma, \delta, m_{\rm max}, \chi_{\rm max}\}$. These are reported in Table I and described below. In particular, the quantities γ, $m_{\rm max}$, and $\chi_{\rm max}$ parametrize the population of first-generation (1g) BHs, while the quantities α, β, and δ parametrize the pairing and merger process.

This setup is an excellent test bed for our deep-learning explorations because these simulations are not computationally intensive (thus allowing us to explore different DNN architectures) while at the same time providing a binary BH population that ultimately is not parametric (thus making our approach essential).

A. Simulation design

We generate $N_\lambda = 1000$ sets of population parameters λ using Latin hypercube sampling to efficiently cover the higher-dimensional space [43,78]. With this design, the hyperparameter space (that is, the space of population-level parameters) is split into N_λ equally probable subintervals in each dimension. From the N_λ^6 possible choices, a total of N_λ unique coordinates are drawn such that, for each of the six dimensions, only one of the N_λ subintervals is selected. In general, there are multiple possible realizations of this random draw; we choose to maximize the minimum distance between points, whose values are chosen as the centers of the intervals. Our simulation design is generated with PYDOE[1].

TABLE I. Parameters in our model of hierarchical binary BH merger populations, the symbols we use to identify them, and their bounds. The population parameters $\lambda = \{\alpha, \beta, \gamma, \delta, m_{\rm max}, \chi_{\rm max}\}$ determine the shape of the distribution of first-generation BHs and the properties of the host cluster that can lead to repeated mergers. The bounds on the power-law indices are broad such that the range of training simulations can incorporate more restrictive prior bounds. The source parameters $\theta = \{M_{\rm c}, q, \chi_{\rm eff}, \chi_{\rm p}\}$ are measured by LIGO/Virgo when detecting individual GW events. The bounds on chirp mass encompass the extrema of the GW catalog posteriors and are only used when evaluating the population-level likelihood, as described in Sec. III A.

	Parameter	Symbol	Range
Population, λ	Primary pairing slope	α	$[-10, 10]$
	Secondary pairing slope	β	$[-10, 10]$
	1g mass slope	γ	$[-10, 10]$
	Escape-speed slope	δ	$[-10, 10]$
	Maximum 1g mass	$m_{\rm max}$	$[30\,M_\odot, 100\,M_\odot]$
	Maximum 1g spin	$\chi_{\rm max}$	[0, 1]
Source, θ	Source-frame chirp mass	$M_{\rm c}$	$[5\ M_\odot, 105\,M_\odot]$
	Mass ratio	q	[0, 1]
	Effective aligned spin	$\chi_{\rm eff}$	$[-1, 1]$
	Effective precessing spin	$\chi_{\rm p}$	[0, 2]

B. First-generation black holes

Each cluster is seeded with $N_{\rm BH} = 5000$ BHs (this number is chosen to ensure convergence of the resulting merger distributions; see Ref. [75]). Their masses $m_{\rm 1g}$ are drawn from a simple, truncated, power-law distribution:

$$p(m_{\rm 1g}|\gamma, m_{\rm max}) \propto \begin{cases} m_{\rm 1g}^{\gamma} & \text{if } 5\ M_\odot < m_{\rm 1g} < m_{\rm max} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

with slope $\gamma \in [-10, 10]$, maximum cutoff $m_{\rm max} \in [30, 100]\ M_\odot$, and a fixed lower boundary of $5\ M_\odot$ (thus only describing black holes and not neutron stars). Pair-instability [79] and pulsation pair-instability supernovae (PISN) [80] prevent the formation of stellar-mass BHs between about 50 and 120 $M_\odot$ [12,13,81]. This prediction is supported by current GW observations, which point to a decrease of the merger rate at those masses [31]. The precise details of the pair-instability mass gap are uncertain and depend on poorly constrained stellar-physics processes such as the nuclear reaction rates [12,13,82], rotation [13,83], accretion [84–88], winds [89,90], envelope retention [91–93], and dredge-up episodes [94]. We thus allow for a broad range of values of $m_{\rm max}$ and aim to infer it from the GW data.

The BH spin directions are drawn from an isotropic distribution, as expected in dynamical environments. The dimensionless spin magnitudes are uniformly within $[0, \chi_{\rm max}]$, where the maximum natal spin is $\chi_{\rm max} \in (0, 1)$. The largest spin formed from stellar collapse is uncertain and difficult to model; see Refs. [95,96]. The spin model we use for first-generation BHs is therefore not necessarily physically well-motivated, but it is used for illustrative purposes.

C. Repeated mergers

At each hyperparameter coordinate, we simulate $N_{\rm cl} = 500$ clusters with escape speeds $v_{\rm esc}$ drawn according to

$$p(v_{\rm esc}|\delta) \propto \begin{cases} v_{\rm esc}^{\delta} & 0\ {\rm km\,s^{-1}} < v_{\rm esc} < 500\ {\rm km\,s^{-1}} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\delta \in [-10, 10]$. Large positive (negative) values of δ give escape-speed distributions skewed towards the maximum (minimum) value of $v_{\rm esc}$. For context, the escape

[1] pythonhosted.org/pyDOE.

speed of a typical globular cluster is 10–100 $\mathrm{km\,s^{-1}}$, while those of nuclear star clusters are up to an order of magnitude larger [97–99]; we take an upper limit of 500 $\mathrm{km\,s^{-1}}$ to accommodate these larger escape speeds. Cases with large, negative δ essentially describe isolated stellar evolution, where repeated mergers do not take place (though we always assume isotropically distributed spins, not partial alignment as expected in isolated binary evolution [14–16]). On the other hand, $\delta = 0$ corresponds to a flat $v_{\rm esc}$ distribution, favoring all environments equally.

The key ingredient in our populations is the presence of so-called "higher-generation" BHs that have undergone multiple mergers due to remnant retention in the host cluster. We form circular binary systems by selectively pairing cluster members according to

$$p(m_1|\alpha) \propto m_1^{\alpha}, \qquad p(m_2|\beta, m_1) \propto \begin{cases} m_2^{\beta} & \text{if } m_2 \leq m_1 \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $m_1 \geq m_2$ are the component BH masses. As for the other power-law indices, we again take $\alpha, \beta \in [-10, 10]$; this broad range is taken in each case so that the simulated populations encompass the prior bounds used later in our statistical inference of Sec. III A. One by one, BH pairs are drawn from the collection according to Eq. (3), and the properties of their merger remnants are estimated (assuming a uniform sampling of the orbital phase) with the implementation of Ref. [100], which collects various numerical relativity fitting formulas [72,73,101–106]. Upon merging, the remnant BHs receive a gravitational recoil [107,108]. If the magnitude $v_{\rm kick}$ of this kick velocity exceeds the escape speed of the host cluster, i.e., $v_{\rm kick} > v_{\rm esc}$, the remnant BH is removed and does not merge again. Otherwise, it remains inside the cluster and can undergo subsequent mergers. The estimated remnant mass and spin magnitude are retained, while the spin directions are resampled isotropically. This pairing, merger, and ejection procedure is iterated until a single BH remains.

For each merger, we record the source parameters $\theta \coloneqq \{M_{\rm c}, q, \chi_{\rm eff}, \chi_{\rm p}\}$. In particular, $M_{\rm c} = (m_1 m_2)^{3/5}/(m_1 + m_2)^{1/5}$ is the chirp mass, $q = m_2/m_1 \leq 1$ is the mass ratio, $\chi_{\rm eff} \in [-1, 1]$ is the effective aligned spin [109], and $\chi_{\rm p} \in [0, 2]$ is a suitable parameter encoding the dominant effect of orbital-plane precession; for the latter, we use the augmented definition of Ref. [110], which consistently averages over the precessional motion including effects from both component spins. While this definition of $\chi_{\rm p}$ is still a frequency-dependent quantity over the inspiral timescale, recent work has shown that the influence of the GW reference frequency at the population level is currently subdominant compared to measurement errors [36]. In the simulated populations, we measure $\chi_{\rm p}$ at the reference frequency of 20 Hz.

Additionally, we record whether each merger is that of two first-generation BHs (1g + 1g) that produces a second-generation (2g) remnant, a first- and second-generation BH (1g + 2g), or two second-generation BHs (2g + 2g), or whether it contains a component BH of higher generation (> 2g). From these, we compute the fraction of mergers in each generation: $f_{\rm 1g+1g}$, $f_{\rm 1g+2g}$, $f_{\rm 2g+2g}$, and $f_{\rm >2g} = 1 - f_{\rm 1g+1g} - f_{\rm 1g+2g} - f_{\rm 2g+2g}$.

D. Cosmic placement

The distribution of sources is assumed to be isotropic over the sky, inclination, and polarization angle. We do not infer the redshift distribution of BH binaries but consider it fixed, i.e., independent of the hyperparameters λ. Each merger is placed at a redshift z according to a distribution that is uniform in comoving volume $V_{\rm c}$ and source-frame time, i.e.,

$$p(z) \propto \frac{1}{1+z} \frac{dV_{\rm c}}{dz}. \tag{4}$$

An immediate generalization of this work would include taking into account the longer assembly times of higher-generation binaries (e.g., Ref. [111]) via their redshift distribution. This can be implemented with an additional hyperparameter and will be tackled in future work.

Ostensibly, $z \in (0, \infty)$, but in practice, there is a detector-dependent horizon, $z_{\max}$, beyond which binary BH mergers are not observable. To find a conservative $z_{\max}$, we consider a series of binaries with aligned maximal spins, equal masses, and optimal orientation with respect to a single detector (overhead and face on). These are the loudest sources for a given total mass and redshift. We compute signal-to-noise ratios (SNRs) as described in Sec. III C and find that the entire mass range becomes subthreshold above an upper bound $z_{\max} = 2.3$, which we thus take as the maximum of the redshift distribution (in agreement with Appendix E of Ref. [30]).

E. Resulting populations

The above prescription allows us to transform a simple phenomenological description of first-generation BH populations into a complex numerical distribution containing hierarchical mergers. The combined set of hyperparameters $\lambda = \{\alpha, \beta, \gamma, \delta, m_{\max}, \chi_{\max}\}$ are very interdependent, and changes in their values cause large variations in the distributions of source parameters $\theta = \{M_{\rm c}, q, \chi_{\rm eff}, \chi_{\rm p}\}$. The total set of simulated events is $\{\{\theta_j^i\}_{j=1}^{N_{\rm h}(\lambda^i)}\}_{i=1}^{N_\lambda}$, where $N_{\rm h}(\lambda^i)$ is the number of mergers occurring in the simulation with hyperparameter coordinate λ^i. The total number of mergers occurring at a given hyperparameter coordinate depends on the distribution of escape speeds, determined by δ. For the numerical setup adopted here, it ranges from $\min_\lambda N_{\rm h}(\lambda) = N_{\rm cl} N_{\rm BH}/2 = 1.25 \times 10^6$ (when each remnant

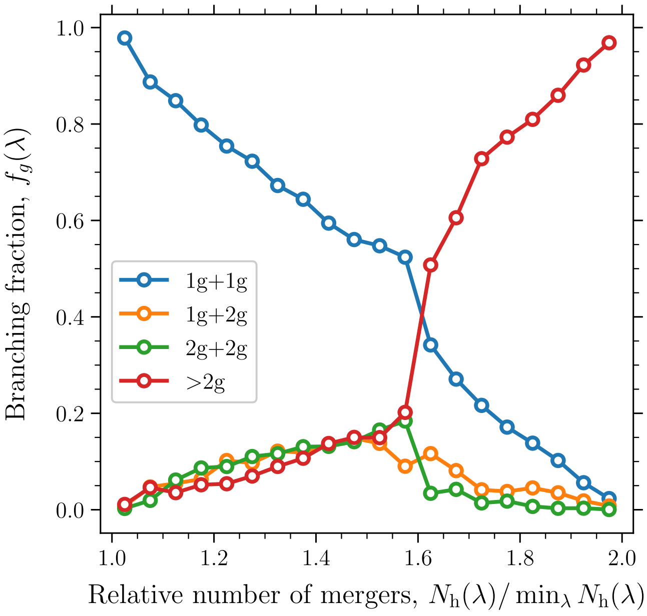

FIG. 2. Fraction of mergers in our simulations from each binary generation as a function of the total number of mergers. The simulations are separated into bins equally spaced in the total number of mergers and the bin-averaged branching fraction of each binary generation—1g + 1g (blue), 1g + 2g (orange), 2g + 2g (green), and higher generations (red)—is plotted. At the lower (upper) end, simulations are dominated by mergers between first- (higher-) generation BHs.

BH is ejected, so only first-generation mergers occur) to $\max_\lambda N_\mathrm{h}(\lambda) = N_\mathrm{cl} N_\mathrm{BH} - 1 \approx 2.5 \times 10^6$ (when BHs are repeatedly paired with the same single retained remnant, i.e., a cluster catastrophe), and the upper range is populated by simulations with larger numbers of repeated mergers. This is demonstrated in Fig. 2, where we plot the branching fractions of different merger generations as a function of the total number of mergers. Four representative cases among the set of $N_\lambda = 1000$ simulations we performed are illustrated in Fig. 3 and labeled based on the qualitative properties of the resulting source distributions: broad masses, narrow mass ratio, broad mass ratio, and repeated mergers.

If clusters are preferentially formed with larger escape speeds, many remnants are retained and proceed to take part in hierarchical mergers, leading to multiple modes in the mass distributions. This is the case for the red curves (repeated mergers) in Fig. 3, where $\delta = 5.1$. Since the sharp initial mass function (IMF) ($\gamma = 5.5$) forms first-generation BHs with masses that are all very close to the maximum $m_\mathrm{max} = 70\ M_\odot$, hierarchical mergers appear as distinct peaks in the mass distributions. The first generation of mergers has $m_1 \approx m_2 \approx m_\mathrm{max}$, giving $M_\mathrm{c} \approx 50\ M_\odot$. Cross-generational mergers also occur. For example, there is a 1g + 2g peak; the peak does not occur at $q = 0.5$ because a fraction $1 - \epsilon \approx 5\%$ of mass is lost via GWs [112] such that second-generation BHs have mass of approximately $2\epsilon m_\mathrm{max}$, implying $q = 1/(2\epsilon) \approx 0.53$ and $M_\mathrm{c} \approx 80\ M_\odot$. Similarly, for a 1g + 3g merger, one has $q \approx 1/[\epsilon(2\epsilon+1)] \approx 0.36$, which explains the third peak observed in the red curves of Fig. 3.

When more first-generation BHs are born with large spins, set by χ_max, fewer second-generation mergers occur due to the larger imparted recoils [74]. On the other hand,

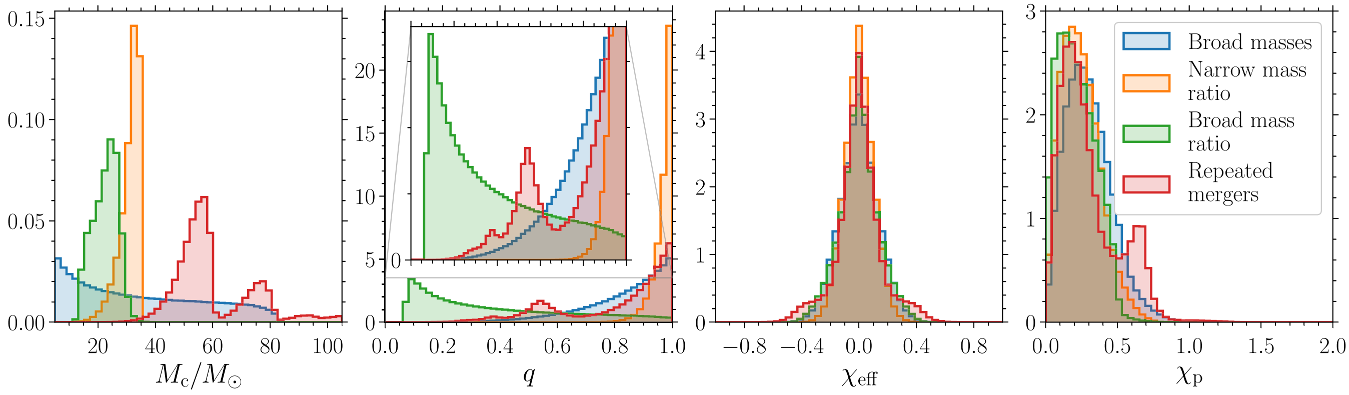

FIG. 3. Example marginal distributions of chirp mass M_c, mass ratio q, effective aligned spin χ_eff, and our precession parameter χ_p for different population parameters $\lambda = \{\alpha, \beta, \gamma, \delta, m_\mathrm{max}, \chi_\mathrm{max}\}$. We select four of our simulations to illustrate different features of the resulting binary BH distributions. In blue, we show broad masses, $\lambda = \{-1.7, 1.7, -0.5, -3.4, 96\ M_\odot, 0.57\}$; this set of hyperparameters results in a large range of binary BH masses due to a high maximum first-generation mass, broad mass function, and broad binary pairing probabilities. In orange, we show narrow mass ratios, $\lambda = \{-8.8, 8.3, 6.8, -4.1, 40\ M_\odot, 0.43\}$; binaries are preferentially selected with equal component masses due to pairing probabilities that favor the lightest primary BHs and heaviest secondary BHs. In green, we show broad mass ratios, $\lambda = \{9.2, -9.8, -0.5, -4.0, 74\ M_\odot, 0.50\}$; the pairing slopes produce binaries with the heaviest primaries and lightest secondaries, resulting in an extended range of mass ratios peaking at lower values. In red, we show repeated mergers, $\lambda = \{4.1, 3.1, 5.5, 5.1, 70\ M_\odot, 0.37\}$; clusters are preferentially generated with large escape speeds, boosting the presence of repeated mergers, which appear as multiple narrow peaks in the mass distributions. The lower maximum natal spin causes a narrow peak around $\chi_\mathrm{eff} = 0$; the occurrence of repeated mergers extends the tails of the χ_eff distribution and creates a secondary peak in the χ_p distribution.

if natal spins are small and repeated mergers do occur, the distribution of effective spins features a sharp peak at $\chi_{\rm eff} = 0$ from first-generation mergers as well as extended tails from high-generation mergers, as is the case for the red curve in the third column of Fig. 3. The $\chi_{\rm eff}$ distributions are always symmetric about zero due to the assumption of spin isotropy. For the 1g + 2g populations, the 2g BH spin is approximately 0.7 [17] and, because in this case $\chi_{\rm max} = 0.37$, is typically higher than the spin of the 1g BH. In this limit, one has $\chi_{\rm p} \approx \sqrt{0.7^2 - 4\chi_{\rm eff}^2} \approx 0.7$ [76], thus explaining the secondary peak in the $\chi_{\rm p}$ distribution.

Whether higher-generation BHs pair with other BHs of equal generation or form cross-generational binaries (e.g., 1g + 2g) depends on the pairing slopes α and β. If α, β, $\gamma \approx 0$, then the first-generation mass distribution is broad, and binary components are selected with uniform probabilities leading to an extended range of mass ratios, as seen in the blue "broad masses" curves of Fig. 3. If α, $\beta \gg 0$ (α, $\beta \ll 0$), the heaviest (lightest) BHs are preferentially selected for both binary components, leading to a heavier (lighter) first generation of approximately equal-mass binaries. If $\alpha \ll 0$ and $\beta \gg 0$ ($\alpha \gg 0$ and $\beta \ll 0$), then the lightest (heaviest) primaries and heaviest (lightest) secondaries are paired, leading to mass-ratio distributions that are sharply peaked at unity (broad and peaked at lower values), as seen in the orange "narrow mass ratio" (green "broad mass ratio") curve of Fig. 3. In the case of narrow mass ratios, given the maximum first-generation mass $m_{\rm max} \approx 40\ M_\odot$ and since $q \approx 1$, the chirp mass peak is located at $M_{\rm c} \approx 35\ M_\odot$.

III. DEEP-LEARNING-ENHANCED POPULATION INFERENCE

Although challenging to treat, a set of highly degenerate hyperparameters makes our simplified population model indicative of realistic applications where GW events are modeled using, e.g., stellar population-synthesis codes. As shown below, deep learning is the ideal tool for such a complex scenario. First, we review the key ingredients that enter hierarchical Bayesian inference to recover the hyperparameters of a population model given GW data from a catalog of mergers (Sec. III A). We then present our method to model the population prior (Sec. III B) and selection effects (Sec. III C) using deep learning. We use similar techniques to model the branching fractions between different merger generations (Sec. III D).

A. Hierarchical Bayesian inference

Given observational data $d = \{d_n\}_{n=1}^{N_{\rm obs}}$ of $N_{\rm obs}$ independent GW events and a population model $p_{\rm pop}$, our goal is to infer the parameters λ governing the shape of the underlying distribution of binary BH source parameters ϑ. The distribution of predicted sources is given by $dN/d\vartheta = N p_{\rm pop}(\vartheta|\lambda)$, where $\int {\rm p}_{\rm pop}(\vartheta|\lambda){\rm d}\vartheta = 1$ is normalized over the entire domain of source parameters. Here, we have separated the parameters that determine the shape λ and overall scale N of the population. Note also that $\vartheta \supset \theta$ is a superset of the source parameters $\theta = \{M_{\rm c}, q, \chi_{\rm eff}, \chi_{\rm p}\}$ we wish to model and additionally contains, e.g., redshift, sky location, inclination, etc. In our case, the extra parameters do not depend on the population-level parameters, such that $p_{\rm pop}(\vartheta|\lambda) = p_{\rm pop}(\theta|\lambda)p_{\rm pop}(\bar\theta)$, where $\bar\theta \coloneqq \vartheta\backslash\theta$, and the normalization over ϑ implies that $\int p_{\rm pop}(\theta|\lambda){\rm d}\theta = \int p_{\rm pop}(\bar\theta){\rm d}\bar\theta = 1$.

1. Selection effects

We wish to infer the observable population of merging BHs from the small subset that we have observed. This requires modeling detector selection effects. The expected number of detectable sources for a given population model is

$$N_{\rm det}(\lambda) \coloneqq \iint \frac{d^2N}{d\vartheta dt} P_{\rm det}(\vartheta, t){\rm d}\vartheta{\rm d}t, \qquad (5)$$

where $P_{\rm det}(\vartheta, t)$ is the probability that a binary BH with source parameters ϑ is detectable at an observation time t (this is a probability and not a probability density, as distinguished by the use of capital P). We describe the calculation of $P_{\rm det}$ in Sec. III C 1. The detection efficiency—i.e., the fraction of detectable events given the population model—is given by

$$\sigma(\lambda) \coloneqq \frac{N_{\rm det}(\lambda)}{N(\lambda)} = \frac{1}{T}\iint p_{\rm pop}(\vartheta|\lambda)P_{\rm det}(\vartheta, t){\rm d}\vartheta{\rm d}t, \quad (6)$$

where we have assumed equally likely arrival times of GW signals at the detectors over the observing period of duration T. The integral over time indicates that we must account for the detector duty cycle and change in sensitivity over observing epochs. We approximate the sensitivity as constant within each observation period: the combined first and second run (O1 + O2) and the third run (O3). The corresponding two-detector observing periods are $T_{\rm O1+O2} \approx 166$ days [3,113] and $T_{\rm O3} \approx 275$ days [4,6], respectively. With this approximation, the time integral reduces to the weighted average [26]

$$\sigma(\lambda) = \sum_r \frac{T_r}{T}\int p_{\rm pop}(\vartheta|\lambda)P_{\rm det}(\vartheta, r){\rm d}\vartheta, \qquad (7)$$

where $r \in \{{\rm O1+O2}, {\rm O3}\}$ indicates the observing run and corresponding instrument sensitivity, and $T = T_{\rm O1+O2} + T_{\rm O3}$ is the total observing time.

2. Population likelihood

Including selection effects, the likelihood of the GW data $\{d_n\}_{n=1}^{N_{\rm obs}}$ given the parameters λ of our population model is (see, e.g., Refs. [25,26])

$$\mathcal{L}(d|\lambda, N) = e^{-N_{\rm det}(\lambda)} \prod_{n=1}^{N_{\rm obs}} \int \frac{dN}{d\vartheta_n} \mathcal{L}(d_n|\vartheta_n) {\rm d}\vartheta_n. \quad (8)$$

The single-event likelihoods $\mathcal{L}(d_n|\vartheta_n)$ may be rewritten using Bayes's theorem as $\mathcal{L}(d_n|\vartheta_n) \propto p(\vartheta_n|d_n)/\pi(\vartheta_n)$, where $p(\vartheta_n|d_n)$ is the posterior on the source parameters for the nth event as inferred by parameter estimation, and $\pi(\vartheta_n)$ is the prior used in that analysis (which may differ event to event). Using Bayes's theorem again, the posterior distribution of population parameters is given by

$$p(\lambda|d) \propto \pi(\lambda) \prod_{n=1}^{N_{\rm obs}} \frac{1}{\sigma(\lambda)} \int \frac{p_{\rm pop}(\vartheta_n|\lambda) p(\vartheta_n|d_n)}{\pi(\vartheta_n)} {\rm d}\vartheta_n, \quad (9)$$

where we have marginalized over the rate parameter N with a scale-independent prior $\pi(N) \propto 1/N$ [114] and $\pi(\lambda)$ is the prior over the remaining shape parameters. The priors on the parameters $m_{\rm max}$ and $\chi_{\rm max}$ are uniform over the ranges listed in Table I. The priors of the power-law indices α, β, γ, and δ are uniform over $[-8, 8]$; these prior bounds lie within the training data range, and we checked that resulting posteriors are robust to more stringent constraints.

3. Factorization of the observed volume

While the integrals in Eqs. (7) and (9) are formally defined over the entire domain of source parameters, in practice, they can be safely performed within the observable volume $V_{\rm h} \coloneqq \{\vartheta : z < z_{\rm max} = 2.3\}$, beyond which the detection probability is zero, as discussed in Sec. II D. Even if $p_{\rm pop}$ models the binary BH population outside of $V_{\rm h}$, as it appears in both the numerator and through $\sigma(\lambda)$ in the denominator of Eq. (9), one can safely assume that $\int_{V_{\rm h}} p_{\rm pop}(\vartheta|\lambda) {\rm d}\vartheta = 1$.

Since it will be useful in Sec. III B, we define the observed volume $V_p \coloneqq \{\vartheta : p(\vartheta|d_n) > 0 \ \forall \ n\} \subset V_{\rm h}$ as the subset of the observable volume beyond which all single-event posteriors $p(\vartheta_n|d_n)$ vanish. For the events considered in this work (see Sec. III A 4 and the Appendix), we find that $V_{\rm p}$ corresponds to $M_{\rm c} \in [5, 105]\ M_\odot$, while for q, $\chi_{\rm eff}$, and $\chi_{\rm p}$, we maintain their natural bounded domains ([0, 1], [−1, 1], and [0, 2], respectively). It will also be useful to define the population prior of our modeled parameters θ normalized over the observed volume,

$$p'_{\rm pop}(\theta|\lambda) \coloneqq \frac{p_{\rm pop}(\theta|\lambda)}{\int_{V_p} p_{\rm pop}(\theta|\lambda){\rm d}\theta}. \quad (10)$$

Since $p_{\rm pop}$ is normalized over $V_{\rm h}$, we can write this extra normalization factor as

$$\int_{V_p} p_{\rm pop}(\theta|\lambda){\rm d}\theta = \frac{N_p(\lambda)}{N_{\rm h}(\lambda)} \leq 1, \quad (11)$$

where $N_{\rm h}(\lambda)$ is the number of mergers occurring within the horizon volume $V_{\rm h}$ and $N_{\rm p}(\lambda)$ is the number of mergers occurring within the subset $V_{\rm p} \subset V_{\rm h}$, given population parameters λ. For convenience, we refactor this term into the detection efficiency by defining the selection function

$$\sigma'(\lambda) \coloneqq \frac{N_{\rm h}(\lambda)}{N_{\rm p}(\lambda)} \sigma(\lambda). \quad (12)$$

By separating the source parameters and noting that our population prior and the parameter estimation prior over the unmodeled parameters are equal, i.e., $p_{\rm pop}(\bar{\theta})/\pi(\bar{\theta}) \equiv 1$, the hyperposterior in Eq. (9) may be written as

$$p(\lambda|d) \propto \pi(\lambda) \prod_{n=1}^{N_{\rm obs}} \int_{V_{\rm p}} \frac{p'_{\rm pop}(\theta_n|\lambda) p(\theta_n, \bar{\theta}_n|d_n)}{\sigma'(\lambda)\pi(\theta_n|\bar{\theta}_n)} {\rm d}\theta_n {\rm d}\bar{\theta}_n. \quad (13)$$

Since the parameter estimation prior is placed on detector-frame masses, we must convert the prior on detector-frame chirp mass $M_{\rm c}^{\rm det}$ to the source frame. In particular, we have $\pi(\theta|z) = \pi(M_{\rm c}^{\rm det}, q, \chi_{\rm eff}, \chi_{\rm p}|z) |\partial M_{\rm c}^{\rm det}/\partial M_{\rm c}|$. Since the Jacobian is $|\partial M_{\rm c}^{\rm det}/\partial M_{\rm c}| = 1 + z$ and the prior on detector-frame masses is independent of the prior on redshift for the parameter estimation results we use below, we have $\pi(\theta|z) = \pi(M_{\rm c}^{\rm det}, q, \chi_{\rm eff}, \chi_{\rm p})(1 + z)$.

4. Event samples

Given discrete samples $\{\{\vartheta_{n,k}\}_{k=1}^{S_n} \sim p(\vartheta_n|d_n)\}_{n=1}^{N_{\rm obs}}$ from the individual event posteriors, where S_n is the number of samples in the posterior for the nth event, and since these samples lie, by definition, within the posterior volume $V_{\rm p}$, Eq. (13) can be evaluated with the Monte Carlo summation

$$p(\lambda|d) \propto \pi(\lambda) \prod_{n=1}^{N_{\rm obs}} \frac{1}{\sigma'(\lambda) S_n} \sum_{k=1}^{S_n} \frac{p'_{\rm pop}(\theta_{n,k}|\lambda)}{\pi(\theta_{n,k}|z_{n,k})}. \quad (14)$$

For each event, we draw prior samples for $\{M_{\rm c}^{\rm det}, q, \chi_{\rm eff}, \chi_{\rm p}\}$ and compute $\pi(M_{\rm c}^{\rm det}, q, \chi_{\rm eff}, \chi_{\rm p})$ using Gaussian kernel density estimates (KDEs) as implemented in SCIPY [115], modified to enforce reflective boundary conditions [116]. Each KDE is then evaluated on the single-event posterior samples. Equation (14) is sampled using DYNESTY [117] and BILBY [118].

We select the confident binary BH detections made during the first (O1), second (O2), and third (O3) observing runs, employing a threshold minimum false alarm rate (FAR) of less than 1 yr^{-1} across all search analyses. This results in a catalog of $N_{\mathrm{obs}} = 69$ binary BH events. For the events in O1 and O2, we use samples[2] from the reanalysis of Ref. [119] because the precession parameter χ_{p} depends on the azimuthal spin angles whose posteriors were not released in GWTC-1 [3]. For the events in O3, we take the posterior samples combining analyses with waveforms including both precession and higher-order modes as provided by the GWTC-2[3] [4], GWTC-2.1[4] [5] (`PrecessingSpinIMRHM`), and GWTC-3[5] [6] (`C01:Mixed`) data releases. We list all of the events that enter our analysis in the Appendix.

B. Population model

The results of our simulations are lists of binary BH mergers, characterized by source parameters $\theta = \{M_{\mathrm{c}}, q, \chi_{\mathrm{eff}}, \chi_{\mathrm{p}}\}$, at each of the $N_\lambda = 1000$ population parameter coordinates, $\lambda = \{\alpha, \beta, \gamma, \delta, m_{\max}, \chi_{\max}\}$. Our approach to modeling the resulting population distribution $p'_{\mathrm{pop}}(\theta|\lambda)$ employs a combination of probability density estimation and regression algorithms.

1. Density estimation

At each of the hyperparameter locations $\{\lambda^i\}_{i=1}^{N_\lambda}$, we evaluate the conditional population density $p'_{\mathrm{pop}}(\theta|\lambda^i)$ with a Gaussian KDE. To efficiently evaluate p'_{pop} with sufficient resolution in the four-dimensional space of source parameters, we use a version of the convolution-based implementation in KDEPY [120], which we modify to enforce the parameter limits (Table I) with reflective boundary conditions [116]. With this method, density estimations of multivariate data with millions of samples evaluated on millions of points take seconds on a standard, off-the-shelf machine, compared to hours with standard KDE routines (the evaluation points must, however, lie on a linearly spaced Cartesian grid that bounds the data extrema). Each dimension is individually scaled with bandwidths determined by the Improved Sheather Jones (ISJ) plug-in selection rule [121,122]. The ISJ algorithm does not make the assumption of normality on the underlying distribution and, as such, is more robust when determining optimal bandwidths for non-Gaussian multimodal distributions. We evaluate each of the N_λ KDEs on a linearly spaced Cartesian grid, including the parameter bounds, with 21 points in each axis.

2. Regression with a deep neural network

Elucidating the scale of the regression problem, there are $21^4 \approx 2 \times 10^5$ KDE evaluations estimating $p'_{\mathrm{pop}}(\theta|\lambda)$ over the combined ten-dimensional vectors of source and population parameters (θ, λ) at each of the $N_\lambda = 1000$ hyperparameter locations. While the KDEs approximate the N_λ functions $\{\theta \mapsto p'_{\mathrm{pop}}(\theta|\lambda^i)\}_{i=1}^{N_\lambda}$, we must also interpolate over the population parameters to find an accurate mapping $(\theta, \lambda) \mapsto p'_{\mathrm{pop}}(\theta|\lambda)$.

To achieve this result, we make use of a fully connected DNN implemented with Google's TENSORFLOW deep-learning library [123]. The network performs a regression of the KDE values of p'_{pop} over the space of (θ, λ) coordinates. As a preprocessing step, we normalize all coordinates (θ, λ) to a unit hypercube using the limits given in Table I, while the values of p'_{pop} are similarly scaled between zero and their maximum. The input layer has $\dim(\theta) + \dim(\lambda) = 10$ neurons, while the output layer has one neuron with enforced non-negativity corresponding to the predicted value of the probability density. Between the input and output layers, the network architecture consists of five hidden layers, each with 128 neurons. We summarize the network architecture in Table II. The number of parameters in a given layer is given by the number of weights (equal to the product of the number of neurons with that of the preceding layer) plus the number of biases (equal to the number of neurons).

We use randomized leaky rectified linear units (RReLUs) [124] in each layer. This modifies the standard rectified linear unit (ReLU) activation function, given by $\mathrm{ReLU}(x) \coloneqq \max(0, x)$, in two ways. First, leaky ReLU activation functions are maps $x \mapsto \max(0, x) + \min(0, ax)$, where $a \in [l, u]$ is a parameter fixed to a small number; i.e., the positive region is linear with unit slope, while the negative region is linear with slope a. Second, the

TABLE II. Architecture of the DNN that emulates the simulated populations by predicting the conditional density $p'_{\mathrm{pop}}(\theta|\lambda)$ of the source parameters θ given population-level parameters λ. Each row represents a single layer of the network and lists the number of neurons in the layer, the activation function of those neurons (RReLU for the hidden layers and absolute value for the final output), and the corresponding number of free parameters.

Layer	Neurons	Activation	Parameters
Input	10	⋯	0
Dense 1	128	RReLU	1408
Dense 2	128	RReLU	16,512
Dense 3	128	RReLU	16,512
Dense 4	128	RReLU	16,512
Dense 5	128	RReLU	16,512
Output	1	Absolute value	129
		Total	67,585

[2] dcc.ligo.org/LIGO-P2000193/public.
[3] gw-openscience.org/GWTC-2.
[4] gw-openscience.org/GWTC-2.1.
[5] gw-openscience.org/GWTC-3.

randomized leaky variant RReLU samples a uniformly in $[l, u]$ during training and fixes $a = (l + u)/2$ when making predictions (we keep the default values of $l = 1/8$ and $u = 1/3$ [124]). Empirically, we find that, among other ReLU variants and nonlinear activations, RReLU gives the best predictive performance while reducing overfitting to the training data.

We split the $N_\lambda = 1000$ simulations into a training data set of 900 runs and a validation set of 100 runs. The validation sample is unseen by the training process except to assess the network performance. The training data input to the network, which is randomly shuffled at each iteration, thus consists of approximately 1.75×10^8 values of the ten-dimensional vector (θ, λ) and the corresponding KDE estimates of $p'_{\rm pop}(\theta|\lambda)$. The network is trained using the Adam optimizer [125], the mean absolute error (MAE) loss function, a learning rate of 10^{-4}, and batch size equal to 0.01% of the total number of training data points. Training is performed for 10^4 epochs on an NVIDIA A100 Tensor Core GPU, taking about four days. With this setup, the number of training epochs is sufficient to ensure convergence of the MAE; the average gradient of the (smoothed) validation MAE over the penultimate 100 epochs is less than 0.1% that of the first 100.

When making predictions with the trained NN, the values are first rescaled from the unit interval to the probability density parameter space. While the predictions are approximately normalized, the network does not enforce unit normalization. Therefore, we estimate normalization factors $\int p'_{\rm pop}(\theta|\lambda){\rm d}\theta$ by numerically integrating the predicted distributions.

In Fig. 4, we summarize the training procedure and predictive performance of our NN population model. The convergence of the MAE loss function for the training and validation samples is plotted in the top panel. The NN fits slightly better to the training data—the validation MAE being, on average, about 1.2 times larger—but there is no significant overfitting. In the bottom panel of Fig. 4, we quantify this statement by comparing the predictive accuracy of the trained population model using the Hellinger distance [126], a metric $d_{\rm H}$ over the space of probability densities that measures the "distance" between two distributions. For two probability densities p and q, it is given by

$$d_{\rm H}(p,q)^2 = \frac{1}{2}\int [\sqrt{p(x)} - \sqrt{q(x)}]^2 {\rm d}x. \qquad (15)$$

Here, $d_{\rm H}$ has the desirable properties of being symmetric and bounded in [0, 1], with $d_{\rm H}(p,q) = 0$ only when $p \equiv q$ and $d_{\rm H}(p,q) = 1$ when p and q have disjoint supports (see Appendix C of Ref. [127] for a physics-oriented summary of the properties of the Hellinger distance). For each of our simulations, we compute the distance between the KDE evaluation and the NN prediction for the probability density. While the mild overfitting presents itself as a small number of outliers at larger values of $d_{\rm H}$ in the validation distribution, both the training and validation subsets have median values of approximately 0.06 and are consistent with each other.

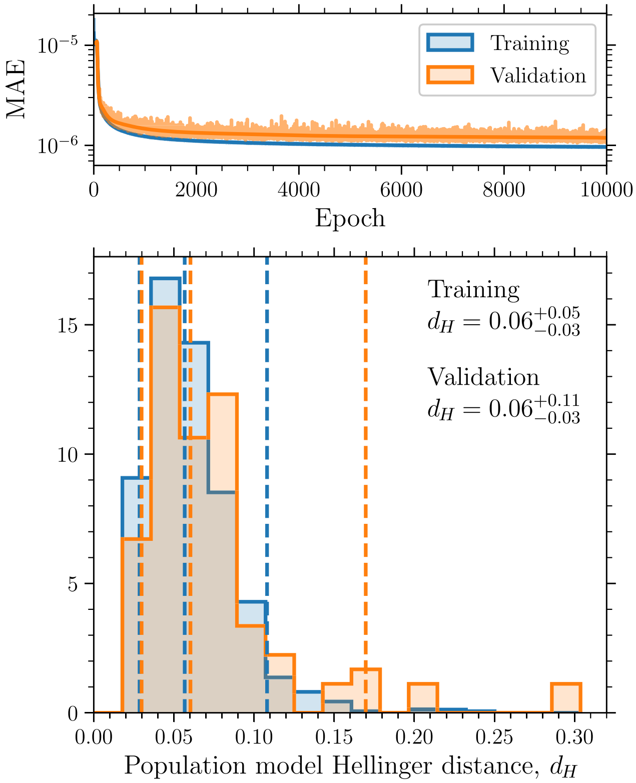

FIG. 4. Top panel: loss functions versus epoch for the training (blue) and validation (orange) data of the population density NN $p'_{\rm pop}(\theta|\lambda)$. Smoothed versions are overplotted in bold. Bottom panel: distribution across all simulations of the Hellinger distances d_H between the true KDE evaluations of $p'_{\rm pop}(\theta|\lambda)$ and those predicted by the NN. The medians and 90% intervals of d_H are plotted as vertical dashed lines and listed explicitly.

In Fig. 5, we illustrate example predictions from our deep-learned population model. For a given set of population-level parameters λ, the NN predicts the value of the joint four-dimensional probability density over the source parameters $\theta = \{M_{\rm c}, q, \chi_{\rm eff}, \chi_{\rm p}\}$. For three validation simulations, we plot the predicted values of $p'_{\rm pop}(\theta|\lambda)$ (solid lines) along with the true KDE evaluations for comparison (circle markers), numerically marginalizing to one-dimensional distributions for the purpose of visualization.

The first example (red) has good predictive accuracy, with $d_H = 0.10$. Here, we use the same distribution labeled "repeated mergers" in Fig. 3, with parameters $\alpha = 4.1$, $\beta = 3.1$, $\gamma = 5.5$, $\delta = 5.1$, $m_{\rm max} = 70\ M_\odot$, and $\chi_{\rm max} = 0.37$. Here, the larger escape velocities and sharp mass function and pairing probabilities lead to distinct

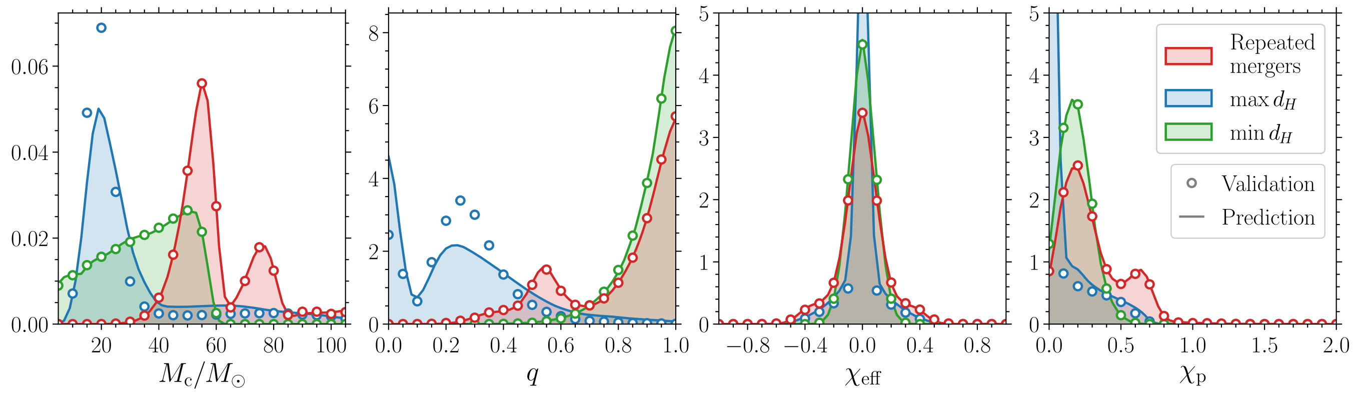

FIG. 5. True KDE evaluations (circle markers) of the population density $p'_{\rm pop}(\theta|\lambda)$ compared against the NN population model predictions (solid lines) for three validation simulations. The full four-dimensional distributions are marginalized to each one-dimensional event-level parameter (left to right: chirp mass $M_{\rm c}$, mass ratio q, effective aligned spin $\chi_{\rm eff}$, and effective precessing spin $\chi_{\rm p}$) for the purpose of visualization. In blue, we show the validation simulation that has the worst predictive accuracy, with a Hellinger distance of $d_H = 0.30$ and population-level parameters $\alpha = 6.3$, $\beta = -7.3$, $\gamma = 1.8$, $\delta = 8.7$, $m_{\rm max} = 46\ M_\odot$, and $\chi_{\rm max} = 0.01$. In green, we show the validation simulation with the smallest Hellinger distance $d_H = 0.02$ and $\alpha = -1.9$, $\beta = 5.2$, $\gamma = 0.5$, $\delta = -9.4$, $m_{\rm max} = 67\ M_\odot$, $\chi_{\rm max} = 0.35$. In red, we show a validation simulation (as in Fig. 3) with $d_{\rm H} = 0.10$ and whose distribution contains distinct features due to repeated mergers.

peaks due to higher-generational mergers. Even though the Hellinger distance of this simulation is greater than the median value, the one-dimensional marginal predictions present excellent matches to the true validation data, accurately capturing all sharp features.

The second case (max d_H, in blue) is a very conservative bound on the performance of our NN, taking the validation simulation with the largest value of the Hellinger distance $d_H = 0.30$ (i.e., that with the worst predictive accuracy). The population parameters are $\alpha = 6.3$, $\beta = -7.3$, $\gamma = 1.8$, $\delta = 8.7$, $m_{\rm max} = 46\ M_\odot$, and $\chi_{\rm max} = 0.01$. While the distributions of the spin parameters $\chi_{\rm eff}$ and $\chi_{\rm p}$ are still fairly well captured, the predictions in the mass distributions suffer from larger errors, though the main features have been learned. The small value of the maximum natal spin $\chi_{\rm max} = 0.01$ leads to sharply peak effective spins $\chi_{\rm eff}, \chi_{\rm p} \approx 0$, while the pairing process generates smaller mass ratios. We stress that this is the worst case among the entire validation set and a rather extreme outlier (cf. Fig. 4). Figure 5 presents the marginalized distributions, while the model predicts the full four-dimensional density, meaning errors over the full source parameter are propagated to the one-dimensional marginals.

The third case (min d_H, in green) represents the best predictive accuracy of our population model, with $d_H = 0.02$. In this validation simulation, the hyperparameters are $\alpha = -1.9$, $\beta = 5.2$, $\gamma = 0.5$, $\delta = -9.4$, $m_{\rm max} = 67\ M_\odot$, and $\chi_{\rm max} = 0.35$, which produces equal masses and a unimodal distribution in the joint four-dimensional space of source parameters. Unsurprisingly, distributions with a simple feature set like this are easier to learn by our DNN population model.

C. Selection function

1. Detection probability

We assume sources are distributed uniformly in sky location, inclination, and polarization angle. We estimate $P_{\rm det}$ with the widely used single-detector semianalytic approximation of Refs. [128,129], as implemented in the Python package GWDET [130], which relies on computing the SNR of optimally oriented sources with the same intrinsic parameters. This is estimated using PYCBC [131], the IMRPHENOMPV2 waveform approximant [132–134], and noise curves representative of the LIGO detector performance during O1O2[6] and O3[7] [135]. While the analytic marginalization of Refs. [128,129] is, strictly speaking, only valid if one neglects spin precession and higher-order GW modes, the impact of these additional effects is subdominant [50]. Their inclusion requires further modeling, which has also been recently tackled using machine-learning techniques [50,51]; we plan to include these refinements in future versions of our population inference pipeline. We employ a SNR threshold of 8 [136] and thus set $P_{\rm det} = 0$ for all subthreshold binaries.

At each location in the population parameter space $\{\lambda^i\}_{i=1}^{N_\lambda}$, we compute $P_{\rm det}$ for all the binaries in the simulation. They have parameters $\{\vartheta^i_j\}_{j=1}^{N_{\rm h}(\lambda^i)} \sim p_{\rm pop}(\vartheta|\lambda^i)$, allowing us to approximate the refactored detection efficiency of Eq. (12) as

[6] Early high from dcc.ligo.org/LIGO-P1200087-v47/public.
[7] LIGO Livingston from dcc.ligo.org/LIGO-T2000012/public.

$$\sigma'(\lambda^i) = \sum_r \frac{T_r}{T}\left[\frac{1}{N_{\rm p}(\lambda^i)}\sum_{j=1}^{N_{\rm h}(\lambda^i)} P_{\rm det}(\vartheta^i_j, r)\right], \qquad (16)$$

where the term in brackets is the Monte Carlo approximation of the integral in Eq. (7).

2. Regression with a deep neural network

To evaluate the (refactored) detection efficiency at arbitrary values of the population parameters, the function $\sigma'(\lambda)$ must be emulated using the discrete evaluations at λ^i. Here, we also use a DNN with TENSORFLOW [123]. The network architecture consists of an input layer with $\dim(\lambda) = 6$ neurons and a linear output layer with one, corresponding to the predicted value of $\ln\sigma'(\lambda)$. We add three hidden layers with 128 neurons each and RReLU activation. This network architecture is summarized in Table III.

We split the hyperparameter coordinates into the same 90% training and 10% validation simulations as in Sec. III B, though note that the training data here consist only of the hyperparameters λ rather than the joint vector (θ, λ). As a preprocessing stage, we again normalize the input values of $\{\lambda^i\}_{i=1}^{N_\lambda}$ to a unit hypercube and train on the output values of $\{\ln\sigma'(\lambda^i)\}_{i=1}^{N_\lambda}$, which are normalized to the unit interval according to the extrema across the simulations. Predictions are rescaled back to the relevant parameter space. We use Adam optimization [125] with a learning rate of 10^{-3} to minimize the mean squared error (MSE) loss function. At each epoch, the training data are shuffled into batches containing 1% of the training data. We train the network for 2000 epochs on a single Intel Core i5-8365U CPU, which took approximately 4 minutes. The training of this DNN is significantly quicker than that of $p'_{\rm pop}$ since it has input dimensionality $\dim(\lambda) = 6$, corresponding to a much smaller training sample size of 900 and a smaller network architecture.

TABLE III. Architecture of the DNN that predicts the logarithmic selection function $\ln\sigma'(\lambda)$ as a function of the population-level parameters λ. Each row represents a single layer and lists its number of neurons, the activation function used, and the corresponding number of free parameters. All hidden layers employ RReLU nonlinearities.

Layer	Neurons	Activation	Parameters
Input	6	⋯	0
Dense 1	128	RReLU	896
Dense 2	128	RReLU	16,512
Dense 3	128	RReLU	16,512
Output	1	…	129
		Total	34,049

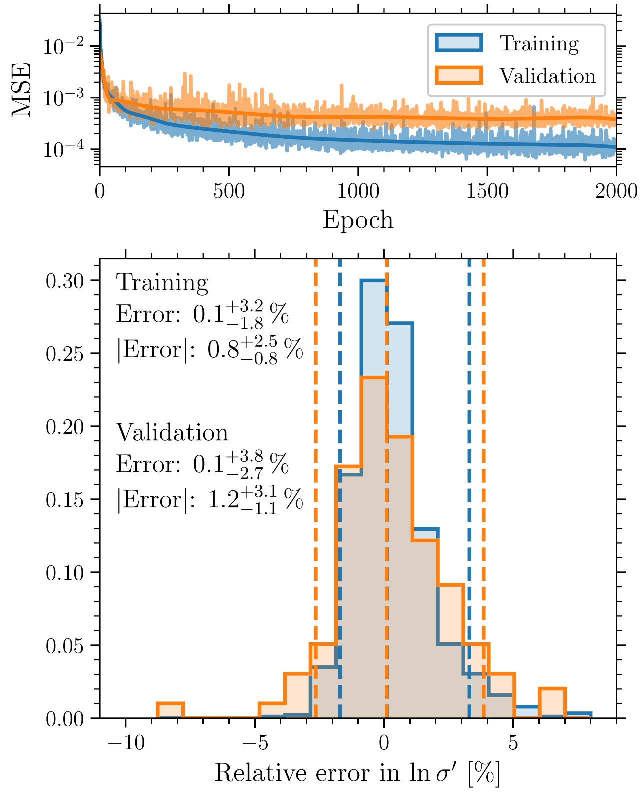

FIG. 6. Top panel: loss function over epochs for the training (blue) and validation (orange) data of the DNN predicting the refactored detection efficiency $\sigma'(\lambda)$. Bottom panel: relative error between the true and predicted values of $\ln\sigma'$. The medians and 90% intervals of the errors are plotted as vertical dashed lines. They are also listed explicitly, as are the magnitudes of the relative errors.

The performance of our DNN to predict $\ln\sigma'$ is reported in Figs. 6 and 7. In the top panel of Fig. 6, we display the convergence of the loss function over the training epochs; the average gradient of the (smoothed) validation MSE over the final 100 epochs is less than or close to 0.5% that over the first 100. While there is some overfitting to the training data, we verify the effect is mild, as follows. In the bottom panel, we display the relative error between the true and DNN-predicted values of $\ln\sigma'$. Since the median and 90% symmetric interval for the validation and training errors are $0.1^{+3.2}_{-1.8}\%$ and $0.1^{+3.8}_{-2.7}\%$, respectively, both are consistent with being centered on and symmetric about zero; i.e., the DNN introduces no systematic biases. The magnitudes of the relative errors for the validation and training sets are consistent with each other and typically less than or close to 5%; the medians and 90% intervals are $0.8^{+2.5}_{-0.8}\%$ and $1.2^{+3.1}_{-1.1}\%$, respectively.

In Fig. 7, we show the dependence of the DNN selection function on each of the hyperparameters for the same three example simulations as in Fig. 5. The true values of $\ln\sigma'(\lambda)$ are shown with circle markers. The predictions of the DNN

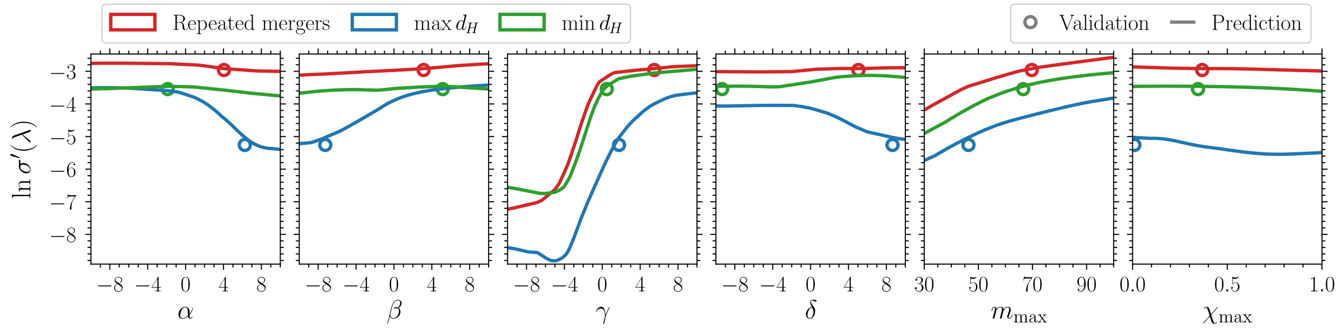

FIG. 7. Example evaluations of the DNN selection function $\ln \sigma'(\lambda)$ at the same three hyperparameter coordinates λ displayed in Fig. 5: a simulation containing repeated mergers (red), and those with the least (blue) and most (green) accurate predictions for the population model DNN. The true value for each simulation is displayed as a circle marker, while predictions made by the DNN are solid lines. In each panel, we vary a single hyperparameter, while the others are fixed to values in the three simulations.

are given by the solid lines, where in each panel we vary a single hyperparameter while keeping the others fixed to the values corresponding to each simulation. For all simulations, σ' is an increasing function of both γ—the power-law index of 1g BHs—and $m_{\max}$—the maximum 1g mass; larger γ implies a greater number of BHs born with masses closer to the maximum $m_{\max}$, while heavier sources emit louder signals and are thus easier to detect (though there is also a compromise with the frequency-dependent—and therefore, mass-dependent—detector sensitivity). The simulation containing repeated mergers (red) consistently features higher values—implying a larger fraction of detectable mergers in the underlying population—due to the larger average binary mass.

The mismatch for the least accurate hyperparameter coordinate of the population model (max d_H, blue) is visible in the offset between truths and predictions. Here, the selection function also depends on α and β, which determine the primary and secondary pairing probabilities. For this simulation, the first-generation mass slope, $\gamma = 1.8$, is quite broad. A wider range of masses implies that a wider range of mass ratios are possible when selecting the BHs in the binary pairing procedure. Higher (lower) values of α (β) lead to higher (lower) primary (secondary) masses and more extreme mass ratios, thus decreasing the detectability. Though repeated mergers occur due to high $\delta = 8.7$, they are preferentially of mixed generations, and therefore, larger δ also leads to lower detectability.

For the validation simulation with the highest population model accuracy (min d_H, green), first-generation masses are broad since $\gamma = 0.5$ and larger since $m_{\max} = 67\ M_\odot$ (compared to $m_{\max} = 46\ M_\odot$ for the max d_H case). Masses are paired equally since $\alpha = -1.9$ selects the lightest primaries and $\beta = 5.2$ selects the heaviest secondaries. The greater prevalence of higher-mass sources with unity mass ratios results in a selection function that is higher (corresponding to increased detectability in the binary BH population) and flatter (with respect to all hyperparameters except γ and $m_{\max}$, as discussed).

D. Merger-generation fractions

As a final demonstration of deep-learning techniques within GW population inference, we train a DNN to infer the branching fractions f_g of the merger generations $g \in \{1\mathrm{g}+1\mathrm{g}, 1\mathrm{g}+2\mathrm{g}, 2\mathrm{g}+2\mathrm{g}, >2\mathrm{g}\}$, as defined in Sec. II C. It is important to note that, unlike the case of branching ratios in mixture population models (e.g., Refs. [49,111,137–139]), these fractions are not hyperparameters themselves but are functions of the model hyperparameters λ. In particular, $f_g(\lambda) = \int p_{\mathrm{pop}}(\theta|\lambda) \mathcal{I}_g(\theta,\lambda) \mathrm{d}\theta$, where $\mathcal{I}_g(\theta,\lambda)$ is a selector function that labels the merger generation, such that $\sum_g f_g(\lambda) \equiv 1$. Our application to the fraction of systems in each hierarchical generation is an example of the more generic problem of constraining formation subchannels that enter a single population.

We use the same training process and network architecture as in Sec. III C 2, with one modification. Since the four branching fractions form a discrete distribution with unit sum, the output layer here has four neurons and employs the activation function $\mathrm{softmax}(\boldsymbol{x})_i \coloneqq \exp(x_i)/\sum_j \exp(x_j)$, where x_i are the components of the input vector $\boldsymbol{x}$. The architecture of this DNN is summarized in Table IV.

In Fig. 8, we plot the converged MSE loss curves. Once more, we assess the accuracy of the DNN predictions against the true generation fractions on the training and validation data sets using the Hellinger distance, which, for discrete probability densities p and q, is given by

$$d_H(p,q)^2 = 1 - \sum_i \sqrt{p_i q_i}. \tag{17}$$

The performances on training and validation subsets are consistent with each other, representing a lack of overfitting. Both have median Hellinger distances of $d_H \approx 0.01$ with $d_H \lesssim 0.1$ for most simulations. The enforced unit summation implies the branching fraction emulator in fact has only three independent outputs despite predicting four

TABLE IV. Structure of the DNN that models the branching fractions f_{1g+1g}, f_{1g+2g}, f_{2g+2g}, and $f_{>2g}$ between the binary merger generations, where, e.g., 1g (2g) denotes a first- (second-) generation component BH. The rows illustrate each layer of the network and report the number of neurons in each, their activation functions (RReLU for the hidden layers and softmax for the output layer), and the number of free parameters.

Layer	Neurons	Activation	Parameters
Input	6	$\cdots$	0
Dense 1	128	RReLU	896
Dense 2	128	RReLU	16,512
Dense 3	128	RReLU	16,512
Output	4	Softmax	516
		Total	34,436

contributions, and in many of our simulations, one or more of the generation labels has zero contribution (e.g., no higher-generation mergers when all remnants are ejected from the host cluster). Both of these considerations produce a tendency for small values of the Hellinger distance, which explains the skew towards $d_H \lesssim 0.01$.

In Fig. 9, we display the dependence of the DNN to predict the branching fractions $f_g(\lambda)$ on the hyperparameters λ for the same validation simulations reported in Figs. 5 and 7. As in Fig. 7, the true values computed from the simulated data are given by circle markers, while predictions made by the DNN are plotted as solid lines where a single hyperparameter is varied while keeping the others fixed. We only display the variation with the power-law indices $\{\alpha, \beta, \gamma, \delta\}$ as we found each $f_g(\lambda)$ to be independent of the maximum first-generation mass $m_{\max}$ and spin $\chi_{\max}$ in these cases. Each branching fraction depends most strongly on the distribution of escape speeds—as determined by the power-law index δ—and the primary binary component pairing probability index α, whereas the indices of the first-generation mass distribution γ and the secondary component pairing β are less impactful.

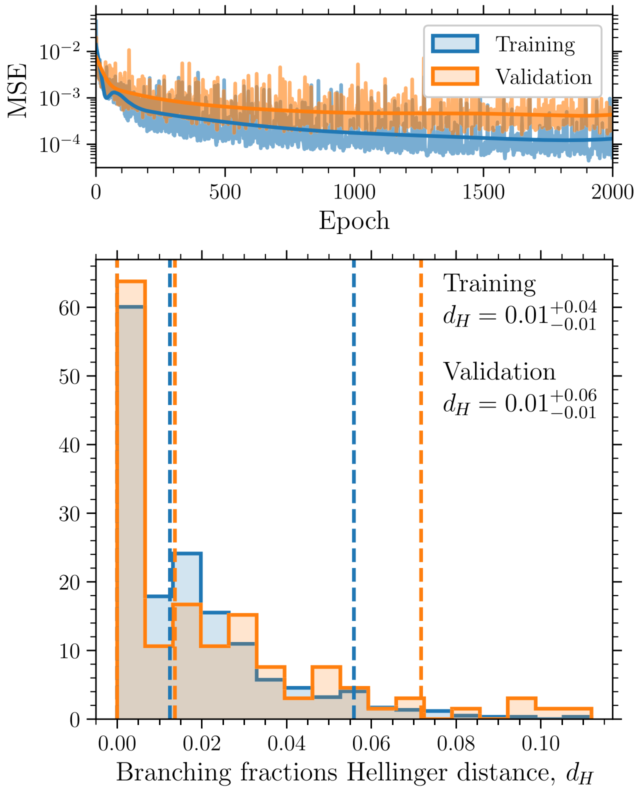

FIG. 8. Top panel: loss functions over training epochs for the training (blue) and validation (orange) data of the DNN predicting the branching fractions f_{1g+1g}, f_{1g+2g}, f_{2g+2g}, and $f_{>2g}$. The actual loss curves are plotted with shading, and smoothed versions are overplotted in bold. Bottom panel: Hellinger distances between the discrete distributions of the true and DNN-predicted merger-generation branching fractions. The medians and 90% confidence intervals are plotted as vertical dashed lines and listed explicitly.

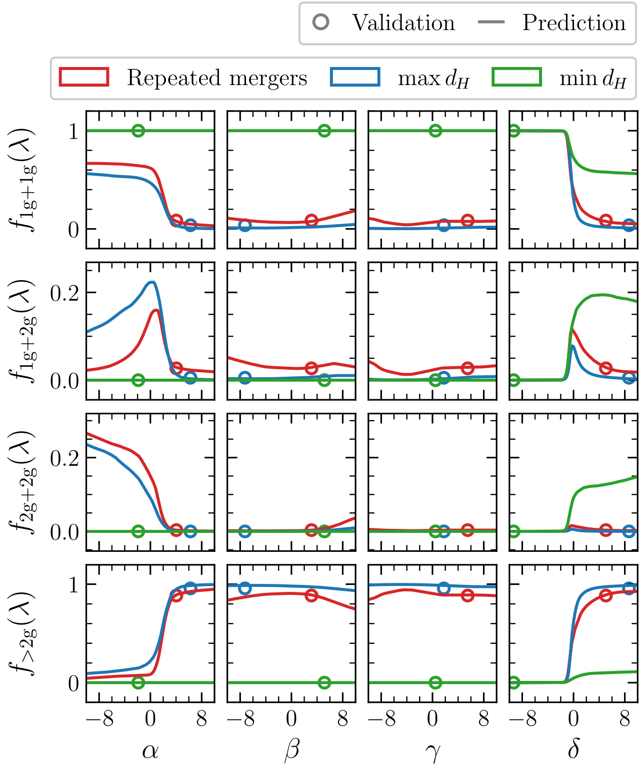

FIG. 9. Example evaluations of the DNN predicting the binary merger-generation branching fractions $f_g(\lambda)$, $g \in \{1g+1g, 1g+2g, 2g+2g, >2g\}$ (from top to bottom rows), as a function of the hyperparameters λ. The results for hyperparameters taken from three illustrative simulations as in Fig. 5—repeated mergers (red), and least and most accurate population predictions (max d_H in blue and min d_H in green, respectively)—are presented. The true values of the generation fractions are plotted as circle markers, whereas DNN predictions are given by solid lines. In each column, a single population-level parameter is varied while keeping the others fixed to those from the simulations.

When $\delta < 0$, the host clusters all have small escape speeds, and therefore, the branching fractions of sources with a remnant BH are close to zero, i.e., $f_{1g+1g} \approx 1$, as seen in the rightmost column of Fig. 9. With a fixed negative δ, as in the case of the green simulation, the branching fractions become independent of the other hyperparameters as no repeated mergers take place. On the other hand, when δ becomes positive, escape speeds are typically larger and repeated mergers can occur, so the contribution to the population from first-generation-only binaries decreases, i.e., $f_{1g+1g} < 1$.

Which binary generation then begins to dominate the population depends on the BH pairing process. When heavier (lighter) primary components form binaries due to a fixed $\alpha > 0$ ($\alpha < 0$), as is the case for the red and blue (green) simulations, > 2g higher-generation (equal second-generation 2g + 2g), binaries preferentially populate the merger distributions. For small positive δ, the contribution from binaries of mixed first and second generations increases, but it is reduced at larger δ in favor of higher-generation mergers. For fixed positive δ (red and blue simulations), larger primary pairing indices α select, with increasingly strong preference, the heaviest remnant BHs in the population to form new binaries, thus increasing the fraction of greater-than-second-generation binaries, i.e., $f_{>2g} \approx 1$, while reducing the prevalence of other generations, as seen in the leftmost column of Fig. 9. The branching fractions are flat for $\alpha < 0$ if, as for the blue simulation, $\gamma < 0$ because all first-generation BHs are typically lighter; therefore, reducing low-mass primary selection bias (i.e., making α less negative) has little effect. In contrast, when $\gamma > 0$ as in the red simulation, first-generation BHs are heavier, and increasing α while keeping β fixed will select heavier primaries relative to the secondaries, therefore favoring 1g + 2g binaries.

IV. VALIDATION WITH MOCK CATALOGS

To test the inference pipeline in the absence of detection biases and single-event measurement uncertainties (equivalent to the limit of large SNRs) and without systematics due to the DNN population, we generate mock GW catalogs by drawing binary BH mergers from our DNN population model $p'_{\rm pop}$. Since for the technical reasons discussed in Sec. III A this distribution is bounded in chirp mass, these draws are inherently taken from that range (listed in Table I). This also means that the selection function constructed in Sec. III C cannot be used in this mock inference; $\sigma'(\lambda)$ is defined over the entire range of source parameters, not just the observed range, and also accounts for the required missing factor between $p'_{\rm pop}(\theta|\lambda)$ and $p_{\rm pop}(\theta|\lambda)$. Including selection effects would require training a different model for the detection efficiency, and thus our tests would include ingredients that do not enter the actual inference of Sec. V. Another technical difficulty is that we model two effective spins, $\chi_{\rm eff}$ and $\chi_{\rm p}$, while the detection probability $P_{\rm det}$, in principle, depends on all six spin degrees of freedom. Creating a mock catalog of observable GW events, i.e., taking samples from the detection-weighted population $P_{\rm det}(\theta)p_{\rm pop}(\theta|\lambda)$, would require assuming an effective lower-dimensional dependence or resampling full spin vectors consistent with the sampled values of $\chi_{\rm eff}$ and $\chi_{\rm p}$ (cf. Ref. [44] for a more in-depth exploration of these issues). However, correctly including spin information in selection biases has a measurable effect at the population level [36].

For testing purposes, we consider the high SNR limit, in which all events are detectable and their source parameters are measured exactly. This corresponds to a selection function $\sigma \equiv 1$ and a single-event likelihood $\mathcal{L}(d_n|\theta) = \delta(\theta - \theta_n)$ for the nth GW event in the catalog. From Eq. (8), the population-level likelihood is thus given by $\mathcal{L}(d|\lambda) \propto \prod_{n=1}^{N_{\rm obs}} p'_{\rm pop}(\theta_n|\lambda)$ (where the statistical details are otherwise equivalent to Sec. III A). We draw $N_{\rm obs} = 50$, 100, 200, 500 events to create increasingly large catalogs (and in going from, e.g., 50 to 100 events, the first 50 are added when increasing the catalog size) with source parameters θ_n ($n = 1, \ldots, N_{\rm obs}$) using rejection sampling of $p'_{\rm pop}$. We repeat the analysis 5 times with new catalogs to assess the impact of population Poisson fluctuations on the inference. To enable a conservative mock catalog test, we fix the true hyperparameters to those of the validation simulation with the lowest predictive accuracy for the DNN population model ($\max d_H$ in Figs. 5, 7, and 9): $\alpha = 6.3$, $\beta = -7.3$, $\gamma = 1.8$, $\delta = 8.7$, $m_{\rm max} = 46\ M_\odot$, and $\chi_{\rm max} = 0.01$.

We present the results of our mock inference runs in Fig. 10. The one- and two-dimensional marginal posterior distributions of the hyperparameters are plotted, where the two-dimensional panels display the 90% contours. The solid black lines denote the true values listed above. In the left panel, we fix the number of observations in the catalog to $N_{\rm obs} = 100$ and perform five independent repetitions of the analysis with five different mock catalogs, given by the different colored curves. Each run is consistent with both the injected hyperparameter values and each other at the 90% level, though there are significant fluctuations between realizations. Recall that the single-event likelihoods neglect measurement errors; relaxing this assumption and including nonzero widths in those posteriors would decrease the overall accuracy of the hyperparameter measurements and thus blend the results from independent realizations to distributions with greater consistency. Increasing the number of observations in the catalog improves the hyperparameter measurement error and reduces the statistical fluctuations between realizations; we take $N_{\rm obs} = 100$ here to approximate the current size of real catalogs [6].

The impact of the growing size of the catalog is illustrated in the right panel of Fig. 10. Here, we choose one particular realization and analyze the catalog as

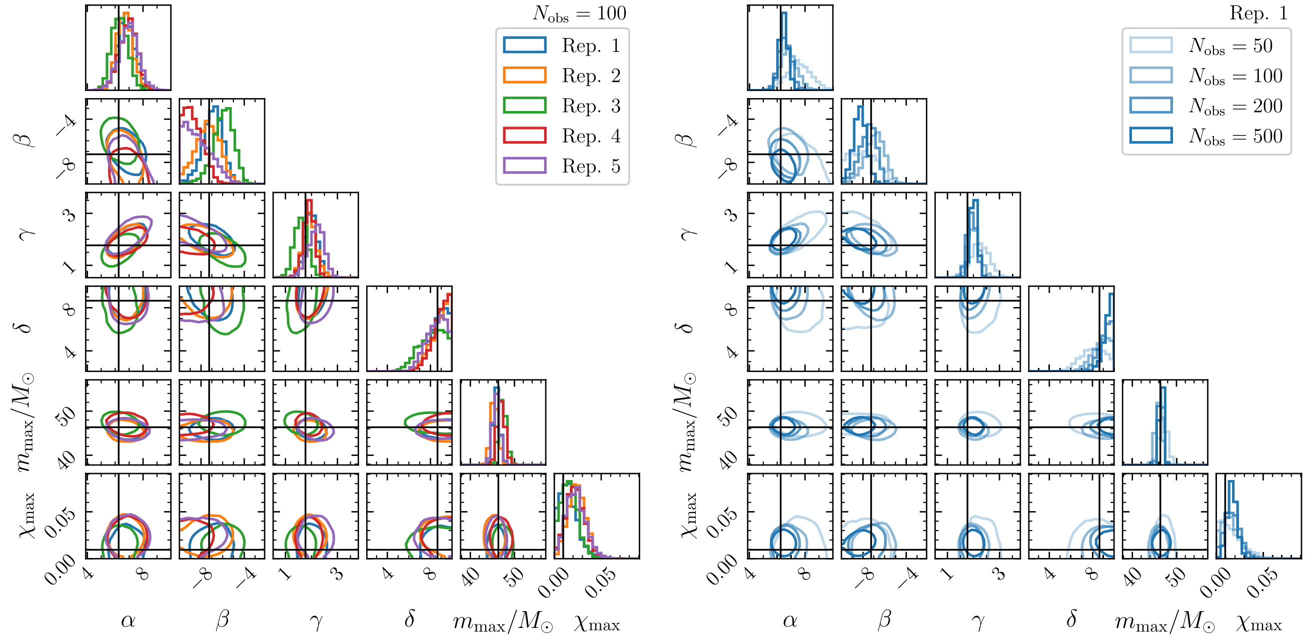

FIG. 10. One- and two-dimensional marginalized posteriors of the population-level parameters $\lambda = \{\alpha, \beta, \gamma, \delta, m_{\rm max}, \chi_{\rm max}\}$, corresponding to the max d_H simulation of Fig. 5, as measured from inference runs without measurement errors and selection biases (corresponding to the high SNR limit), and systematics from the DNN population model by drawing mock GW catalogs directly from $p'_{\rm pop}$. For the joint two-dimensional panels, each contour encloses the 90% credible region for a single analysis. Injected values are marked with black lines. Left panel: number of observations in the catalog fixed to $N_{\rm obs} = 100$ and five independent realizations of the inference with distinct events performed, each represented with a different colored curve (Reps. 1–5). Poisson fluctuations emerge as variations in the Bayesian measurements of the population-level parameters. Right panel: hyperposteriors for mock catalogs drawn for Rep. 1 on the left presented for an increasing number of observed events, $N_{\rm obs} = 50$, 100, 200, 500 (light to dark shading). Larger catalogs break degeneracies between parameters, and the resulting posteriors converge upon the true hyperparameters with tighter constraints.

increasing numbers of events are added incrementally (light- to dark-blue curves). We recover the expected result: The posterior constraints become tighter as $N_{\rm obs}$ increases from 50 to 500 while remaining consistent with the true hyperparameter values at the 90% level. Larger catalog sizes also break degeneracies between parameter pairs, e.g., the β-γ correlations, and remove posterior support in regions far from the truth, e.g., in the column for α.

If the events from the mock catalogs are instead drawn from the simulated populations used as validation samples when training the $p'_{\rm pop}$, one may expect a systematic bias in the recovered hyperposteriors as the number of observations increases due to mismodeling in the trained NN. Indeed, when repeating the above analysis but injecting from simulated validation data while recovering with the NN, we find hyperposteriors that can exclude the injected values at 90% confidence for the lowest accuracy (max d_H) simulation considered above when $N_{\rm obs} \geq 100$. However, this point is a considerable outlier in terms of accuracy (see Fig. 4). For most regions in the hyperparameter space, the mismodeling between injection and recovery remains consistent at 90% confidence. In particular, we verify that this is true for the validation simulation whose hyperparameters are closest to the recovered medians in Sec. V, suggesting our inference on the GWTC-3 catalog below is robust within the measurement uncertainties. While the tests performed here are admittedly limited in scope, they allow us to assess the renormalization and sampling capabilities in the pipeline.

V. POPULATION FROM GWTC-3

In the following, we infer the population properties of the binary BHs in GWTC-3 given our deep-learned population model of hierarchical mergers. In Fig. 11, we present the result of our population inference—the posterior distribution of the hyperparameters λ. Along the diagonal is the one-dimensional marginalization of each hyperparameter, while the other panels display the 50% and 90% confidence intervals of each two-dimensional distribution.

A. Host escape speeds

We begin with the properties of host environments. We consider clusterlike hosts—simply collections of individual BHs that may be paired to form binaries—which are solely characterized by their escape speeds $v_{\rm esc}$. Recall that the simulated clusters are distributed according to a truncated power law $p(v_{\rm esc}|\delta) \propto v_{\rm esc}^{\delta}$, with $0 \ {\rm km\,s^{-1}} < v_{\rm esc}$

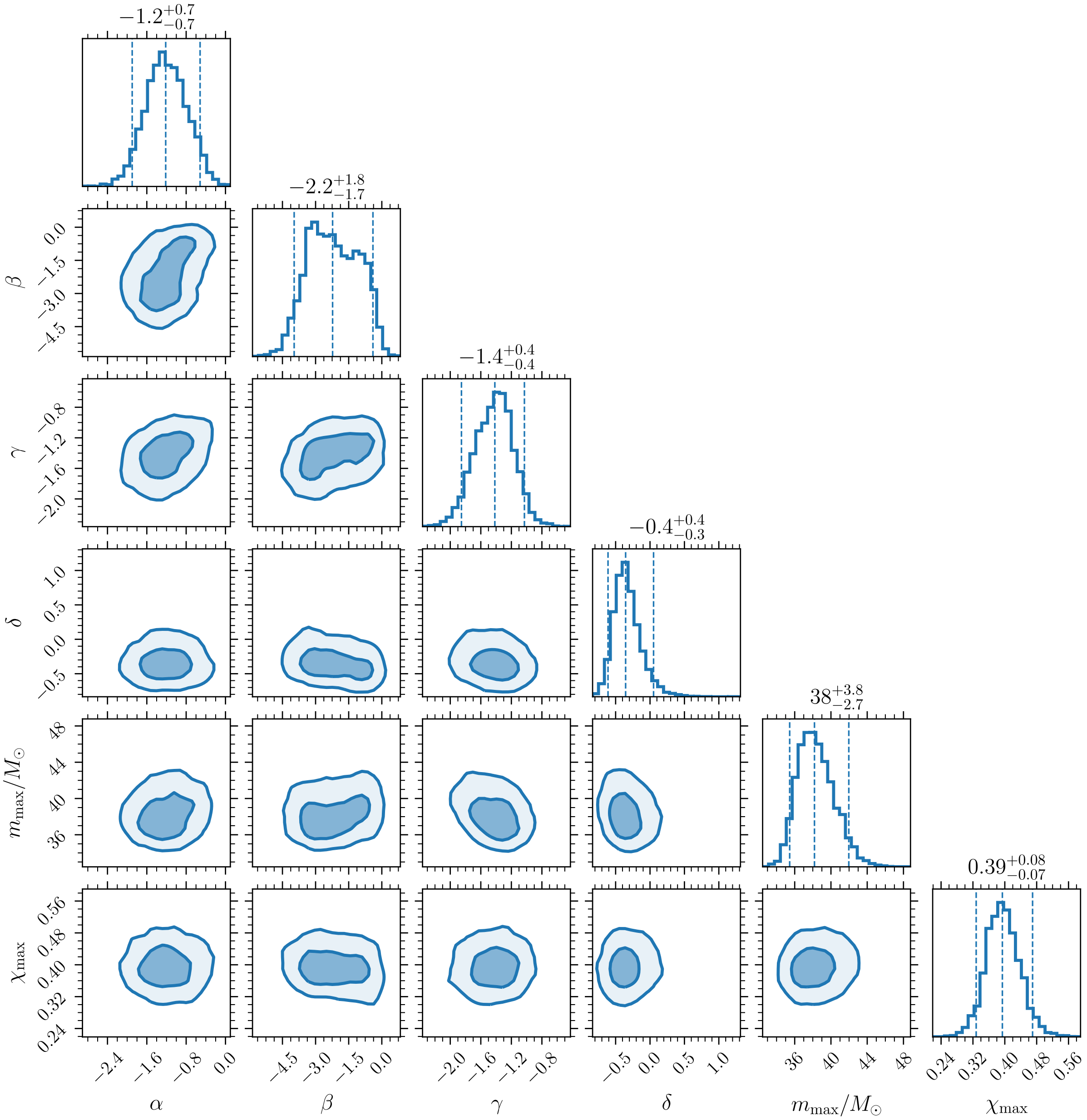

FIG. 11. One- and two-dimensional marginal distributions of the population-level parameters $\lambda = \{\alpha, \beta, \gamma, \delta, m_{\rm max}, \chi_{\rm max}\}$ in our model of hierarchical mergers as measured using the real GW data from the confident (FAR < 1 yr^{-1}) binary BH events through GWTC-3. In each two-dimensional distribution, the contours enclose the 50% (dark shading) and 90% (light shading) confidence regions. The one-dimensional median and symmetric 90% intervals are reported above each diagonal and are plotted as vertical dashed lines in the corresponding panels.

< 500 km s^{-1}. The value of each cluster's escape speed controls whether repeated mergers take place since sources receiving larger gravitational kicks are ejected. Though a power-law distribution is a simplified model, it is indicative of a preference (or lack thereof) towards either edge of the domain.

The marginal distribution of the escape-speed index δ is displayed in the fourth diagonal entry in Fig. 11. Negative (positive) values of δ indicate an escape-speed distribution favoring lower (higher) values, while $\delta = 0$ corresponds to a uniform distribution in $v_{\rm esc}$. We report a median and symmetric 90% interval of $\delta = -0.4^{+0.4}_{-0.3}$, corresponding to an escape-speed distribution biased toward smaller values though consistent with uniformity within the 90% credible bounds.

In Fig. 12, we display the distribution of escape speeds reconstructed from the hyperposterior $p(\delta|d)$. Marginalizing the escape-speed model $p(v_{\rm esc}|\delta)$ over the uncertainty in the hyperparameter δ returns the posterior population distribution (PPD)

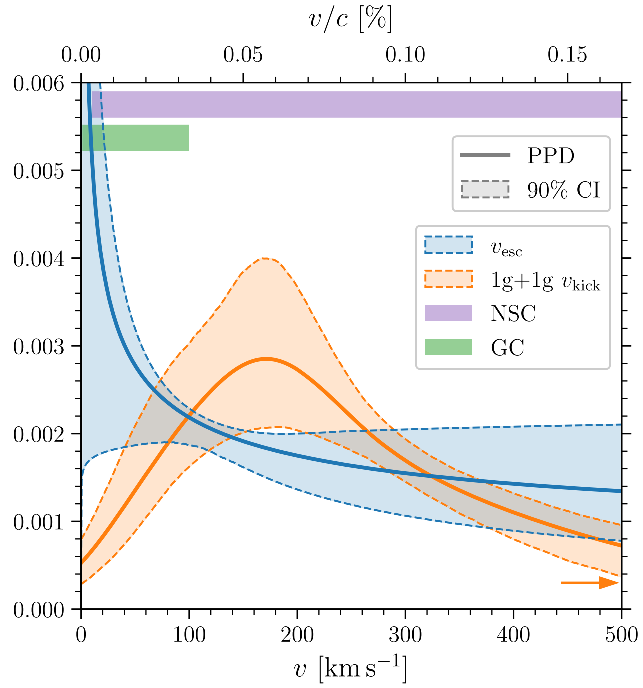

FIG. 12. PPDs of cluster escape velocities $v_{\rm esc}$ (blue), and of gravitational recoils $v_{\rm kick}$ for binaries consisting of two first-generation BHs (orange). The orange distributions are normalized over a range extending beyond the upper $v_{\rm esc}$ limit, as indicated by the arrow. The colored shaded bands contextualize the velocity scale by denoting the typical escape speeds of globular clusters (green) and nuclear star clusters (purple).

$$\mathrm{PPD}(v_{\rm esc}) = \int p(v_{\rm esc}|\delta)p(\delta|d)\mathrm{d}\delta, \tag{18}$$

whereas the posterior uncertainty is displayed interior to the 5% and 95% quantiles of $p(v_{\rm esc}|\delta)$, with $\delta \sim p(\delta|d)$. For context, order-of-magnitude estimates of the escape speeds of globular clusters (GCs; $\lesssim$100 km s^{-1}) and nuclear star clusters (NSCs; $\lesssim$500 km s^{-1}) are shown as horizontal colored bands [97–99]. Additionally, we display the PPD of gravitational kicks, $v_{\rm kick}$, received by the $1\mathrm{g}+1\mathrm{g}$ sources implied by our population model and hyperparameter constraints. Recall that, since the first-generation and binary pairing distributions have parametric forms (see Sec. II), the $1\mathrm{g}+1\mathrm{g}$ distribution does also (i.e., power-law mass distributions, uniform dimensionless spin magnitudes, isotropic spin directions; we present the measurements of the population-level parameters governing this distribution in the following sections). Though GW kicks peak at about 200 km s^{-1}, the distribution of escape speeds features support across the defined range up to $v_{\rm esc} = 500\,\mathrm{km\,s^{-1}}$. For these $1\mathrm{g}+1\mathrm{g}$ sources, we find that $P(v_{\rm kick} < 500\ \mathrm{km\,s^{-1}}) = 0.85^{+0.06}_{-0.08}$ and $P(v_{\rm kick} < v_{\rm esc}) = 0.37^{+0.13}_{-0.12}$, implying that host environments can retain the kicked remnants of a portion of first-generation mergers and support a population of hierarchical BHs.

B. Mass distribution

First-generation BHs—those born in stellar collapse—are drawn according to $p(m_{1\mathrm{g}}|\gamma, m_{\rm max}) \propto m_{1\mathrm{g}}^{\gamma}$, with $5\ M_\odot < m_{1\mathrm{g}} < m_{\rm max}$ providing a mass limit corresponding to the lower (upper) edge of the purported upper (lower) mass gap. We recover $\gamma = -1.4^{+0.4}_{-0.4}$, implying lighter BHs closer in mass to $5\ M_\odot$ (chosen here to conservatively rule out NS and ambiguous source classifications) preferentially populate the underling population. Negative 1g mass power-law exponents are expected, and they reflect the stellar IMF [140].

If first-generation BHs are drawn from this single power-law prescription, we find a first-generation upper mass limit of $m_{\rm max} = 38^{+3.8}_{-2.7}\ M_\odot$. The presence of events in the GW catalog with component masses greater than 50 $M_\odot$ already points to the possibility of hierarchical mergers. Theoretical and simulated estimates of a mass gap location due to PISN typically predict $m_{\rm max} \sim 50\ M_\odot$, but they range within $40\ M_\odot \lesssim m_{\rm max} \lesssim 70\,M_\odot$ (or even higher [141]) due to varying assumptions on key uncertain parameters [12,13]. For comparison, taking the POWER LAW+PEAK GWTC-3 analysis of Ref. [31]—which features a Gaussian peak with mean μ_m to model mass buildup, potentially due to PISN—we find consistency within 90% credible bounds between the inferred $\mu_m = 34^{+2.6}_{-4.0}\ M_\odot$ and $m_{\rm max}$. Note, however, that while $m_{\rm max}$ is a sharp cut specifically characterizing the first-generation BH mass limit, the model of Ref. [31] parametrizes all BHs in a single distribution and μ_m is only the mean of a broadened feature, such that $m_{\rm max}$ and μ_m are not directly equivalent. In our case, BH with masses larger than $m_{\rm max}$ are accommodated with hierarchical mergers.

Here, we point out a key distinction of our modeling procedure: We assume all first-generation component BHs are drawn from a shared distribution (above), and then binary formation is separately modeled with component pairing probabilities $p(m_1|\alpha) \propto m_1^{\alpha}$ and $p(m_2|\beta, m_1) \propto m_2^{\beta}$ ($m_2 < m_1$). This choice differs from, e.g., POWER LAW +PEAK [31], which models each component mass distribution with multiple features superimposed on a power-law distribution. One may be tempted to think that, e.g., the primary mass distribution is equivalent to $p(m_1|\alpha)p(m_1|\gamma, m_{\rm max}) \propto m_1^{\alpha+\gamma}$ (and similarly for secondary masses), however, this applies only to $1\mathrm{g}+1\mathrm{g}$ binaries. Our DNN population model additionally captures the interdependence between binary pairing and remnant retention. In short, the power-law indices parametrizing the distributions in this work are not directly comparable to such models. We infer $\alpha = -1.2^{+0.7}_{-0.7}$ and $\beta = -2.2^{+1.8}_{-1.7}$, such that both component pairing probabilities are bottom heavy with positive power-law indices ruled out at the 90% confidence level.

Having reported the inferred population-level parameters governing the binary BH distributions, we now turn to the implied source-parameter PPDs given by

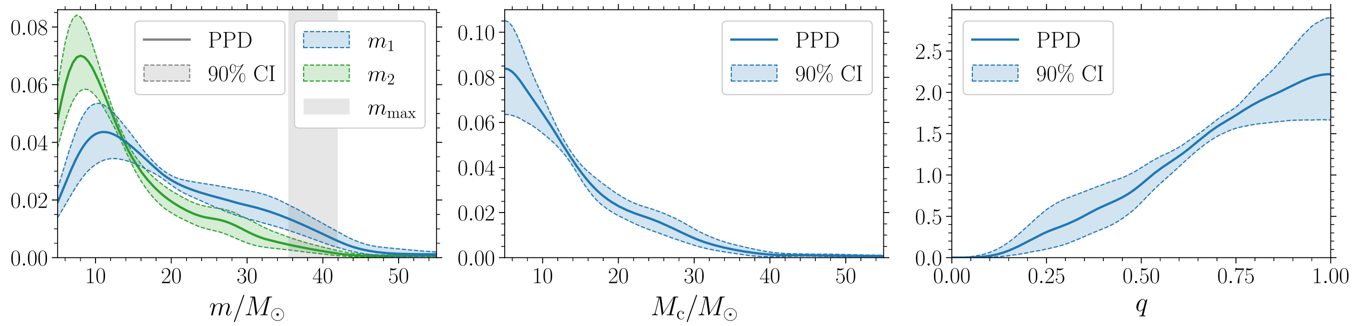

FIG. 13. Astrophysical distributions of the modeled chirp mass $M_{\rm c}$ (middle panel) and mass ratio q (right panel), as well as the implied distributions of primary and secondary masses, m_1 and m_2, respectively (left panel), as determined by our DNN population model and Bayesian analysis of the binary BH merger events in GWTC-3. The solid blue lines represent the PPDs, while the dashed lines enclose the 90% symmetric confidence intervals (shaded). In the left panel, the vertical gray band encloses the 90% confidence interval for the maximum mass of first-generation BHs, $m_{\rm max}$.

$$\mathrm{PPD}(\theta) = \int p'_{\rm pop}(\theta|\lambda)p(\lambda|d)\mathrm{d}\lambda, \tag{19}$$

such that the astrophysical distribution of each source parameter is given by the one-dimensional marginalizations of PPD(θ). In Fig. 13, we present the inferred source distributions of the modeled mass parameters—chirp mass $M_{\rm c}$ and mass ratio q—and the implied distributions of primary and secondary masses, $m_1 = M_{\rm c}(1+q)^{1/5}/q^{3/5}$ and $m_2 = qm_1$, respectively. Each PPD is plotted as a bold solid line, while the symmetric 90% confidence region of each marginal $p'_{\rm pop}(\theta|\lambda)$ with $\lambda \sim p(\lambda|d)$ is represented by shaded bands. The chirp mass distribution peaks at the minimum value 5 $M_\odot$ allowed by our model before an approximately exponential decline, with $M_{\rm c} \lesssim 40\ M_\odot$. Equal-mass binaries are preferred in the underlying population, the mass-ratio distribution having a peak at $q = 1$ but with a broader linear decline down to $q \gtrsim 0.1$.

Substructure is apparent in the distributions of component source masses, corroborating the findings of Refs. [31,142]. Tighter constraints at $M_{\rm c} \approx 13\ M_\odot$ and $q \approx 0.6$ result in a cusp in the primary (secondary) mass distribution around $m_1 \approx 20\ M_\odot$ ($m_2 \approx 12\ M_\odot$) between two features: the peak of the distribution at $m_1 \approx 12\ M_\odot$ ($m_2 \approx 8\ M_\odot$) and a buildup-following decline at the first-generation mass limit $m_{\rm max} \approx 40\ M_\odot$. This suggests two contributions to the mass distribution in the range $20\ M_\odot \lesssim m_1 \lesssim 40\ M_\odot$: (1) first-generation BHs with masses above the peak of the distribution, and (2) higher-generation BHs with masses still smaller than $m_{\rm max}$ but whose parents originally had masses in the peak 10–20 $M_\odot$ region. While high-mass outliers above $m_{\rm max}$ might be considered as clear indicators of repeated mergers, the bottom-heavy nature of the stellar IMF implies that hierarchical mergers may be prominent also for sources with masses below $m_{\rm max}$.

The first-generation and combined component mass distributions are compared in Fig. 14. In purple, we show the reconstructed distribution of first-generation masses,

$$\mathrm{PPD}(m_{\rm 1g}) = \int p(m_{\rm 1g}|\gamma, m_{\rm max})p(\gamma, m_{\rm max}|d)\mathrm{d}\gamma\mathrm{d}m_{\rm max}, \tag{20}$$

and in blue, we show the joint distribution of all primary and secondary masses. The gray shaded band represents the 90% constraint on the mass limit of first-generation BHs, $m_{\rm max}$. Note the logarithmic scale, and that the PPD is a set of expectation values (i.e., means) and, as such, can lie outside the region bounded by given quantiles. Though declining above the first-generation cutoff, the mass distribution features an extended spectrum above $m_{\rm max}$ which cannot result from 1g + 1g mergers. We find that 99% of all BHs have masses less than $59^{+7.8}_{-6.5}\ M_\odot$. The spectrum ultimately abates at $m_1 > 80\ M_\odot$—roughly $2m_{\rm max}$, implying a lack of greater-than-2g mergers with parent components from the upper end of the 1g mass spectrum—and features multiple small-scale modes in the intervening region. These observations again point to hierarchical mergers in the underlying population.

C. Spin distribution

Moving to binary BH spins, recall that the first generation of BHs are modeled with isotropic spins whose dimensionless magnitudes are distributed uniformly up to a maximum $\chi_{\rm max} \in (0, 1)$, representing the maximum natal spin a BH may be born with in stellar collapse. We infer a value $\chi_{\rm max} = 0.39^{+0.08}_{-0.07}$. With limited constraining power in the spin observables, the precise constraints reported here are likely to be very model dependent. We opt for a uniform distribution of 1g spin magnitudes because of the large uncertainties surrounding the spin of compact objects following core collapse (e.g., Refs. [16,95,96,143,144]); this is an area where more accurate observations and more constraining predictions are very much needed. The overall distribution of spins is determined jointly by the first-generation distribution, the binary pairing procedure (as inferred above), the general-relativistic mapping of binary

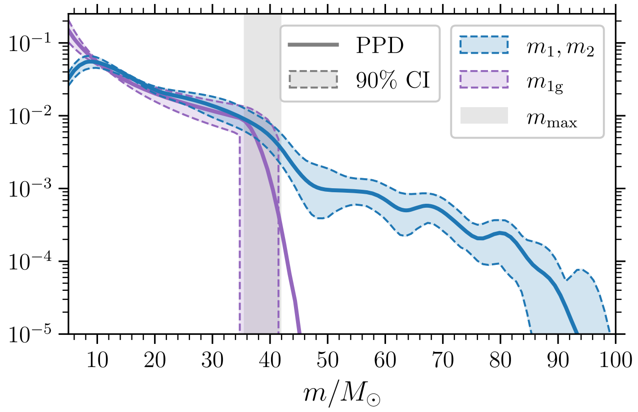

FIG. 14. PPDs (in logarithmic scale) of the first-generation BH masses m_{1g} (purple) and the combined distribution of all components masses m_1, m_2 (blue). The solid lines denote the means, while the dashed lines bound the shaded 90% symmetric confidence regions. The vertical gray band encloses the 90% constraint on the maximum first-generation BH mass, $m_{\max}$.

to remnant properties, and the ejection or retention of merger remnants in host environments. While we account for the dimensionless spin magnitudes of higher-generation binaries in our population modeling, the spin directions are resampled isotropically.

A more solid finding we report is that the spins of $1g + 1g$ binary BHs are limited below the typical dimensionless spin of merger remnants, approximately 0.7 [17]. Hierarchical BHs with much lower spins are extremely rare [77], yet another indication that some higher-generation binary BHs are required to fit the data with our model (cf. Sec. V D). We measure spins using two effective parameters: The effective aligned spin $\chi_{\rm eff}$ measures the binary spin component parallel to the orbital plane [109], and the effective precessing spin $\chi_{\rm p}$ measures the in-plane, two-spin projection [110]. For sources with negligible, misaligned, or (equal-mass) oppositely aligned spins, we have $\chi_{\rm eff} \approx 0$, while large positive (negative) values indicate high aligned (antialigned) spins. Similarly, $\chi_{\rm p} \approx 0$ for spins that are small, aligned with the orbital angular momentum, or oppositely aligned in the orbital plane. Nonzero values of $\chi_{\rm p}$ indicate the presence of spin precession, with $\chi_{\rm p} > 1$ being a region exclusively occupied by binaries with precessing spin contributions from both BH components.

Figure 15 displays the PPDs of these two modeled effective spin parameters. In the left panel, we show the distribution of effective aligned spins $\chi_{\rm eff}$. Here, the assumption of isotropic spins leads to an overly tight constraint. This mismodeling enforces a distribution that is symmetric about and centered on $\chi_{\rm eff} = 0$, in contrast with more generic spin models that infer asymmetric distributions skewed to positive $\chi_{\rm eff}$ [31] (and thus favoring alignment) or those that rule out negative $\chi_{\rm eff}$ [34,35]. However, we find that, typically, $|\chi_{\rm eff}| \lesssim 0.4$, in agreement with the results of Ref. [31] (GAUSSIAN SPIN model); in particular, we report $|\chi_{\rm eff}| < 0.46^{+0.04}_{-0.06}$ for 99% of the population.

On the other hand, the right panel of Fig. 15 shows the distribution of precessing spins measured with $\chi_{\rm p}$, where, unlike Ref. [31], we observe substructure; note that, although they use the earlier $\chi_{\rm p}$ definition of Ref. [145], for the majority of events, the two measurements are indistinguishable [110,146]. We note that, like for $\chi_{\rm eff}$, the uncertainty is likely also underestimated here due to our modeling assumptions. The distribution features two prominent modes. The primary one appears at $\chi \approx 0.2$. A peak at $\chi_{\rm p} > 0$ is determined by the model, given isotropic spin directions (as is the case for all merger generations in

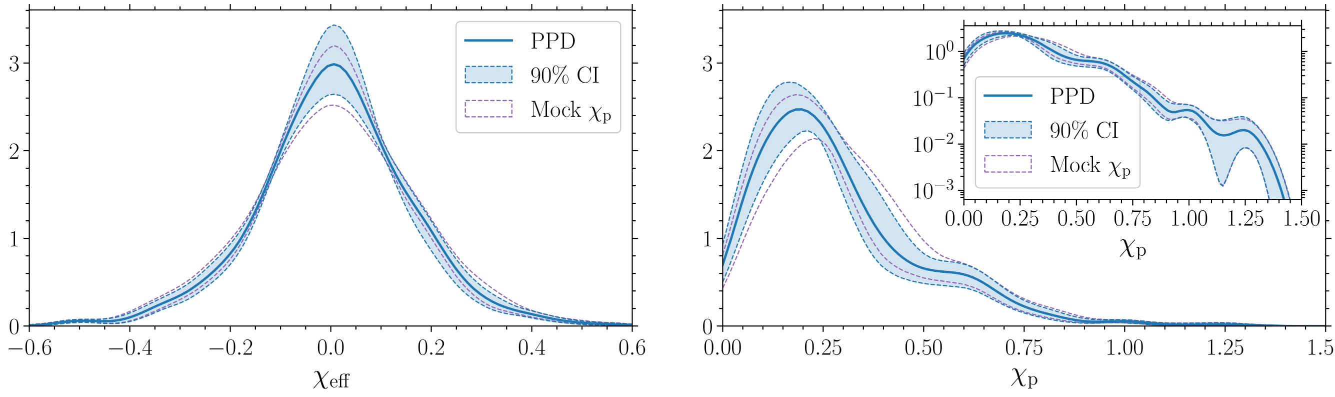

FIG. 15. PPDs of the effective aligned ($\chi_{\rm eff}$, left panel) and precessing ($\chi_{\rm p}$, right panel) spins derived from our DNN population inference of binary BHs in GWTC-3. The means of the distributions are plotted with solid blue lines, and the symmetric 90% confidence intervals are given by the shaded bands bounded by dashed lines. The inset for the $\chi_{\rm p}$ panel shows the same distribution with logarithmic scaling to highlight smaller-scale features. The dashed purple lines bound the 90% confidence region of the distributions that are measured when replacing the true parameter estimation results for $\chi_{\rm p}$ with mock samples from the prior for each event in the catalog.

our model) and uniform nonzero spin magnitudes (as for the first-generation binaries). A single-peaked distribution essentially corresponds to the implied χ_p prior used in parameter estimation analyses [110]. If this feature is astrophysical in origin rather than due to our model choices, however, it may imply that sources with at least moderately misaligned spins—and thus undergoing spin precession—make up a sizable portion of the population. The shape and location of this mode are in broad agreement with the results of Ref. [31]; see their Fig. 16.

However, in contrast to their finding that χ_p measurements can be explained by *either* a narrow distribution with peak $\chi_p \approx 0.2$ *or* a broad distribution centered on $\chi_p = 0$ (which results in multimodality when marginalized over the posterior uncertainty), we find that individual distributions drawn according to the hyperposterior *always* decrease at $\chi_p = 0$, peak at $\chi_p \approx 0.2$, *and* feature a secondary mode typically around $\chi_p \approx 0.6$. While our population model naturally accommodates such multimodal structure, the GAUSSIAN SPIN model employed in Ref. [31] only allows for a single peak and is thus unable to jointly capture the narrow $\chi_p \approx 0.2$ peak in addition to the extended distribution above $\chi_p \gtrsim 0.5$, instead favoring one or the other. Indeed, a single Gaussian distribution cannot fit the distribution of χ_p within the 90% credible bounds.

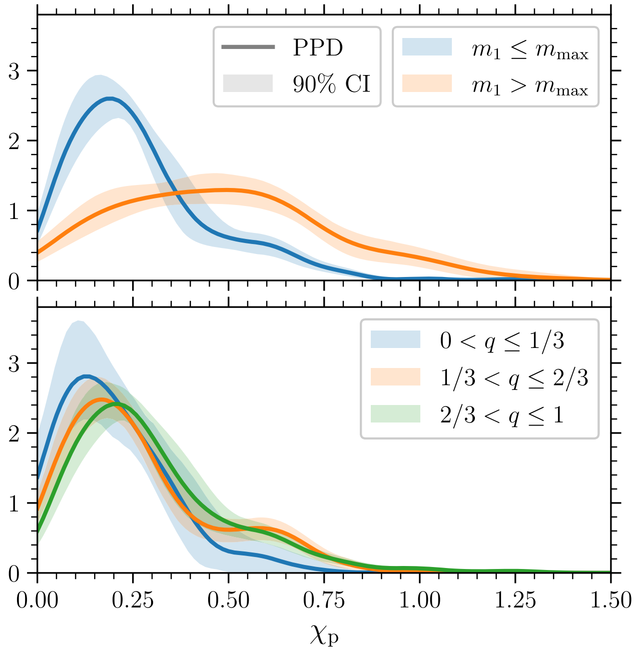

FIG. 16. PPD of the effective precessing spin χ_p as a function of the primary mass m_1 (top panel) and mass ratio q (bottom panel). The mean distributions are given by the solid lines, and the symmetric 90% confidence intervals are given by the shaded bands. In the top panel, we split the χ_p posterior for primary masses below (blue) and above (orange) the maximum first-generation mass $m_{\max}$. In the bottom panel, we split the χ_p posterior for mass ratios $0 < q \leq 1/3$ (blue), $1/3 < q \leq 2/3$ (orange), and $2/3 < q \leq 1$ (green).

The inset in the χ_p panel of Fig. 15 shows the same distribution with logarithmic scaling to highlight smaller-scale features. The distribution falls off above the feature at $\chi_p \approx 0.6$ before a tertiary buildup at $\chi_p \approx 1$ and a final minor mode at $\chi \approx 1.25$, with a large decline in between and eventual declivity beyond. We find minor evidence for a population of sources occupying the exclusive two-spin region $\chi_p > 1$; the 99% quantile lies at $\chi_p = 0.95^{+0.07}_{-0.13}$ while $P(\chi_p > 1) = 0.8^{+0.6}_{-0.4}\%$. There is no support in the population for $\chi_p \gtrsim 1.5$.

Turning to the origins of these spin features, the precessing spin posterior we measure differs from a population prior with uniform spin magnitudes and directions due to the inferred constraint $\chi_{\max} < 1$, leading to a shift towards lower values and, more importantly, the feature at $\chi_p \approx 0.6$, which is not explainable with such a model.

To test whether the posterior constraints are really due to measurements of precession or correlations with other parameters—primarily the best-measured spin χ_{eff} and the mass ratio q—we repeat the hierarchical inference but replace the χ_p posterior for each event in Eq. (13) with samples from the parameter estimation prior. The measured 90% confidence intervals for the effective spin posteriors are shown by the dashed purple lines in Fig. 15. The constraints are qualitatively the same, with only small differences in the 90% credible regions. The purple χ_p distribution favors slightly larger values in the region $\chi_p < 0.6$, suggesting the real precession measurements from GW data offer *some* information beyond the parameter estimation prior, but the differences are minor. The most informative constraints originate from the aforementioned better-measured parameters.

In our single-channel model, a BH with mass above $m_{\max}$ is necessarily a merger remnant. Since merger remnants have large spins, this model requires heavy BHs to have large spins if there are masses above $m_{\max}$ in the catalog and natal spins are small, as inferred above. Our DNN model naturally allows for correlations between parameters, unlike simple phenomenological priors, so we can assess which masses contribute to the $\chi_p \approx 0.6$ feature. In the top panel of Fig. 16, we split the inferred χ_p population posterior into contributions from primary masses $m_1 \leq m_{\max}$, which can be both first- or higher-generation BHs, and definitely higher-generation sources with $m_1 > m_{\max}$. Though the latter, heavier population of sources necessarily has a preference for larger spins, the contribution to the χ_p distribution from sources with $m_1 \leq m_{\max}$ still contains the feature at $\chi_p \approx 0.6$, implying that this inference is not solely driven by the requirement for heavy BHs to have large spins. This is a consequence of the previous conclusion that hierarchical mergers in our

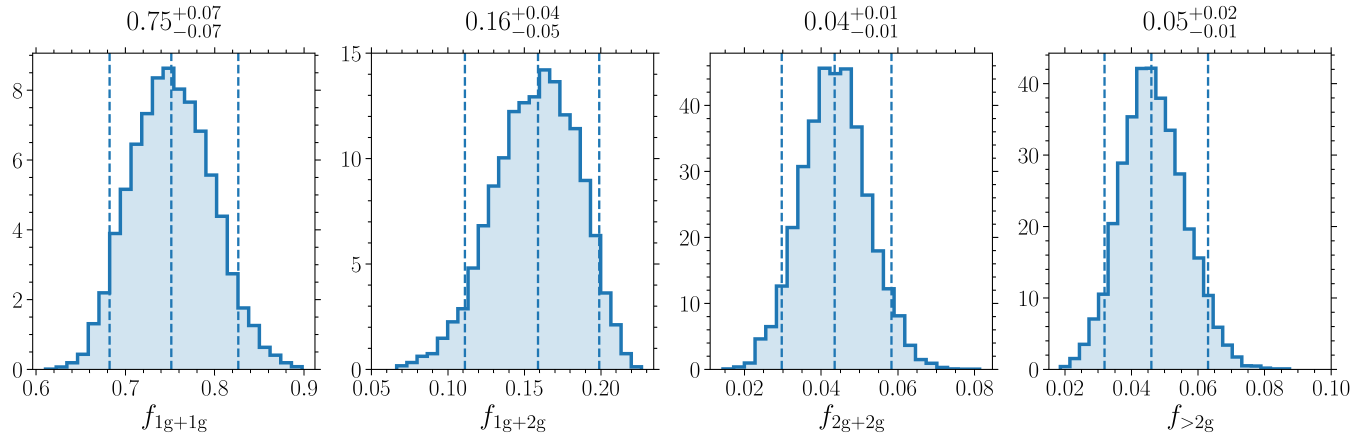

FIG. 17. Distributions of the branching fractions (left to right: f_{1g+1g}, f_{1g+2g}, f_{2g+2g}, $f_{>2g}$) for merger generations in the astrophysical distribution of merging stellar-mass binary BHs, as measured with our deep-learning approach to population inference on the GWTC-3 catalog. The median and symmetric 90% confidence region for each generation fraction is reported above—and plotted as vertical dashed lines within—the corresponding panel.

model also populate the region $m_1 \leq m_{\max}$ due to the bottom-heavy mass function.

In the bottom panel of Fig. 16, we similarly observe the population distribution of χ_p as a function of the mass ratio q. Larger precessing spins are suppressed for more unequal masses $q \leq 1/3$ with the peak lowered to $\chi_p \approx 0.15$, while for mass ratios $q > 2/3$ closer to unity, it increases to the slightly larger value $\chi_p \approx 0.25$. Spins around $\chi_p \approx 0.6$ are present for these mass ratios, but the distinct feature is most prominent for $1/3 < q \leq 2/3$. This region of the parameter space is prominently occupied by mixed-generation mergers, e.g., $1g + 2g$. We verify that this secondary structure is consistent with repeated mergers in our model as follows. Starting with the $1g + 1g$ PPD and binary pairing measurements to compute the distribution of 2g remnant masses and dimensionless spin magnitudes (computed with Ref. [100] as in Sec. II C), the distributions of χ_p for binaries formed either with a 1g BH and a remnant BH (i.e., $1g + 2g$) or two remnant BHs (i.e., $2g + 2g$) both feature peaks at $\chi_p \approx 0.6$. This is because, for $1g + 2g$ sources, the dominant contribution to χ_p is from the primary, which is more likely to be 2g, while for $2g + 2g$ sources, the primary and secondary are more likely to contribute equally such that their average is similar to the $1g + 2g$ case.

D. Merger generations

Given our DNN population model, the observations of the previous sections suggest the presence of hierarchical mergers in the underlying population of merging stellar-mass binary BHs. Taking samples $\lambda \sim p(\lambda|d)$ from the posterior distribution of population parameters in Eq. (14), the corresponding draws from the posterior of merger-generation fractions can be derived as $f_g(\lambda) \sim p(f_g|d)$, where f_g is given by the DNN described in Sec. III D.

Figure 17 presents the posterior distributions of the fractional contribution to the population from each binary merger generation; the medians and 90% symmetric intervals are quoted and indicated as vertical dashed lines. In the underlying distribution, $75^{+7}_{-7}\%$ of sources contain only first-generation BHs ($1g + 1g$), which implies around 25% contain a component that is the remnant of a previous merger, with 90% (99%) one-sided support for $1 - f_{1g+1g} \gtrsim 0.19$ (0.14). Mixed-generation binaries with both a first- and second-generation component make up the second-largest portion of the population, with $f_{1g+2g} = 0.16^{+0.04}_{-0.05}$, while binaries containing two second-generation BHs or any component of even higher generation contribute equally at about the 5% level ($f_{2g+2g} = 0.04^{+0.01}_{-0.01}$ and $f_{>2g} = 0.05^{+0.02}_{-0.01}$, respectively).

Previous studies of older GW catalogs found weak evidence for the presence of hierarchical mergers [147,148]. However, further detections through GWTC-2 brought the addition of events whose properties, including higher masses and mass ratios, hinted at higher-generation origins. Reference [61] presented a population analysis based on a phenomenological model of globular clusters, implying the presence of at least one second-generation BH in the GWTC-2 events with greater than 96% probability, rising to greater than 99.99% when considering their highest Bayes factor model corresponding to an escape speed $v_{\rm esc} \sim 300\ {\rm km\,s^{-1}}$. In this case, they found median relative merger rates of 0.15 and 0.01 when comparing $1g + 2g$ and $2g + 2g$ binaries to the $1g + 1g$ case, respectively, with 99% upper limits of 0.29 and 0.04. Equivalently, in our GWTC-3 analysis, we find broadly consistent relative branching fractions $f_{1g+2g}/f_{1g+1g} = 0.21^{+0.08}_{-0.08}$ and $f_{2g+2g}/f_{1g+1g} = 0.06^{+0.03}_{-0.02}$ (reporting medians and symmetric 90% confidence intervals). Given the disparity of the underlying model assumptions between the two analyses and the addition of new detections in GWTC-3, our results jointly point to the fact that, if admitted in the fitted population, a modest number of binary BHs with hierarchical origin appears necessary to best explain the data.

VI. SUMMARY AND FUTURE PROSPECTS

Our findings were made possible by advances in the treatment of the GW data, in particular, exploiting deep-learning techniques. These aspects are summarized in Secs. VI A and VI B, respectively.

A. Astrophysics summary

We fit current LIGO/Virgo data assuming a population of sources that generalizes current phenomenological functional forms while consistently allowing for the occurrence of hierarchical mergers. Therefore, the crucial feature of our model is the separation of first- and higher-generation merger populations, the latter of which is not phenomenological. This feature allows us to place constraints directly on the properties of those BHs born as stellar remnants in addition to the population as a whole. We summarize our key results as follows (quoting medians and 90% credible regions):

(i) The distribution of escape speeds of environments hosting binary BH mergers is relatively flat, though lower values are preferred; modeled as a power law in the range $0 < v_{\rm esc} < 500\ {\rm km\,s^{-1}}$, the index is $\delta = -0.4^{+0.4}_{-0.3}$. Such environments may retain merger remnants since $37^{+13}_{-12}\%$ of $1{\rm g} + 1{\rm g}$ remnants receive GW recoils $v_{\rm kick} < v_{\rm esc}$.

(ii) When parametrized as a truncated power law (whose minimum is fixed to $5\ M_\odot$), the distribution of first-generation masses has index $\gamma = -1.4^{+0.4}_{-0.4}$ and thus favors lighter BHs. First-generation BHs have an upper mass limit $m_{\rm max} = 38^{+3.8}_{-2.7}\ M_\odot$.

(iii) Negative power-law slopes are recovered for the binary-pairing probability distributions, indicating both components are selected with a preference for lighter BHs, though this preference is stronger for secondaries; the primary (secondary) pairing index is $\alpha = -1.2^{+0.7}_{-0.7}$ ($\beta = -2.2^{+1.8}_{-1.7}$). This finding is inconsistent with uniform binary pairing ($\alpha = \beta = 0$) at the 90% level.

(iv) This results in a primary (secondary) mass distribution that peaks around $m_1 \approx 12\ M_\odot$ ($m_2 \approx 8\ M_\odot$), with a buildup and then decline before the first-generation upper mass limit. Mass ratios peak at unity but extend to $q \gtrsim 0.1$. While 99% of the population has masses less than $59^{+7.8}_{-6.5}\ M_\odot$, there is an extended spectrum beyond the first-generation mass distribution due to repeated mergers.

(v) Assuming a distribution of first-generation BH spins that is isotropic in direction and uniform in magnitude, we find that the maximum spin formed in stellar collapse is $\chi_{\rm max} = 0.39^{+0.08}_{-0.07}$. The distribution of effective aligned spins features support within $|\chi_{\rm eff}| < 0.46^{+0.04}_{-0.06}$. The effective precessing spins are multimodal, with a maximum at $\chi_{\rm p} \approx 0.2$ and a secondary peak due to repeated mergers at $\chi_{\rm p} \approx 0.6$, but they fall off in the two-spin region with less than or about 1% of the distribution at $\chi_{\rm p} > 1$ and vanishing support for $\chi_{\rm p} \gtrsim 1.5$.

(vi) Approximately 25% of binaries in the underlying population contain a higher-generation BH, with 99% one-sided support for a fraction greater than or about 14%.

While we are able to highlight some key insights into the astrophysics of stellar-mass black-hole mergers, we stress that our DNN population model is based on simulations that are simplified by, e.g., employing various power-law parametrizations. This work serves as a test case to demonstrate the efficacy of the modeling procedure by bridging the gap between phenomenological and accurate simulated models.

The complexity of our simulated populations can be increased in various ways. First, we model the spin distributions of first-generation binary BHs as uniform and isotropic, and while we account for the spin magnitude of merger remnants, we continue the assumption of isotropicity in higher-generation mergers. Employing more sophisticated spin magnitude models and adding an additional hyperparameter to control the degree of first-generation spin alignment, we can better capture the behavior of a wider class of host environments, e.g., isolated evolution [15] or the disks of active galactic nuclei [149]. More generally, allowing for contributions from a mixture of distinct formation channels would lead to a more realistic fit. In particular, $\chi_{\rm eff}$ has been shown to favor positive values, which may indicate a significant contribution from isolated binary formation to the merger rate [31], but we only consider a single-channel dynamical formation model. By underestimating the location of the $\chi_{\rm eff}$ distribution, the fraction of hierarchical mergers may be overestimated [150]. The added complexity of multichannel modeling is beyond the scope of this first study, and we aim to address it in future work. Capturing more realistic distributions of higher-generation spins requires both retaining information on postmerger spin directions and characterizing any changes in relative orientation during binary formation.

Second, we model redshifts with a fixed distribution corresponding to a rate of events that is uniform in comoving volume and source-frame time. A simple extension would be to include a parametrized redshift model [114], though this would also increase the dimensionality of the hyperparameter space. We consider the mergers in our simulated populations as an ensemble and do not account for dynamical assembly of hierarchical mergers, i.e., for the fact that a remnant BH can only form a new binary at later times than its parent system [111]. In practice, one should include the effect of time delays between formation and merger and thus more realistically model the merger rate; we leave such explorations to future work.

B. Deep-learning summary

Compared to previous work [43,44], the approach presented here replaces the approximation of simulated binary BH distributions via histograms with Gaussian KDEs, and the emulation of these distributions across both the source- and population-level parameter spaces via GPR with DNNs. While GPR has been shown to be an ineffective approach in higher dimensions [45], alternative deep-learning techniques such as normalizing flows [47] have proven successful [46,48,49].

Rather than training on probability density evaluations (the required number of which scales exponentially with the dimensionality) as in this work, normalizing flows are trained directly on samples from the true distribution (thus scaling linearly with dimensionality), making the latter more effective in high-dimensional spaces. Further, our methodology requires truncating the population model in the unbounded chirp mass parameter in order to generate training data and employ numerical normalization, introducing a refactoring term in the population-level likelihood. This issue may be solved with domain compactification by, e.g., modeling the inverse of the mass scale instead. Normalizing flows also have the advantage of being generative models (i.e., from which new predictive samples can be drawn) that additionally provide density estimation with correctly enforced normalization. However, typically either just the forward (density prediction) or inverse (sample generation) model is efficient to evaluate [47,151]. On the other hand, neither deep-learning approach provides any estimation of modeling uncertainty, whereas GPR does; this is an area where Bayesian deep learning may prove fruitful (see, e.g., Ref. [152] for a hands-on approach).

That said, our DNN framework has some advantages. The separation of density estimation and emulation adds a level of flexibility not otherwise available. For example, outputs of current state-of-the-art stellar-physics codes provide evolutionary tracks that act as proxies for a given contribution to the merger rate, in practice, outputting a set of weighted binary BHs. This can be trivially implemented in our formalism without any modification to the underlying computational framework by including sample weights when fitting the KDEs. The KDE-first approach also allows for sufficient smoothing of the training distributions prior to the learning stage. This distinction is important for simulations with modeling choices that lead to nonphysical numerical discretization of outputs or low sampling densities; e.g., despite employing normalizing flows, Ref. [49] required kernel density resampling to boost the set of simulated BH mergers.

More importantly, in a follow-up study, we will also explore the potential for emulating labeled subpopulations within a given model. Taking the case at hand, one may wish to individually model the distributions of each merger generation rather than the combined distribution as a whole. Each of these are themselves not probability densities due to potential degeneracies in the hyperparameter space and must be modeled correspondingly; e.g., for low escape speeds, the entire population will be contained within the first-generation label while the others will have empty supports. Since each distribution is formed by the same generative process, a single model with multidimensional output should be used to predict the partitioned populations; this requirement can be readily satisfied in our DNN framework with the inclusion of additional output neurons whose activation functions, combined with the overall numerical normalization, constrain predictions to the required unit summation.

VII. CONCLUSIONS

In this work, we have made use of multiple DNNs to tackle several aspects of the inference of GW catalogs. We focused on stellar-mass binary BH mergers, which are target sources of current ground-based GW observatories. In particular, we considered populations of binary BHs containing repeated mergers—systems in which individual BHs may be born not only from stellar collapse (i.e., as first-generation BHs) but as the remnants of (potentially multiple) previous binary mergers (higher-generation BHs).

Starting from simple phenomenological parametrizations of clusterlike progenitor host environments, the mass and spin distributions of first-generation BHs, and binary pairing, we constructed a suite of simulated mergers, using NR fitting formulas to estimate the properties—mass, spin magnitude, and kick—of remnants, self-consistently accounting for their ejection from (or retention in) the host clusters due to gravitational recoil. The resulting hierarchical merger populations are complex and cannot be represented with closed-form expressions, as is precisely the case for more realistic progenitor modeling (e.g., population synthesis simulations). This a textbook case where machine-learning methods can show their full potential.

We trained a high-dimensional DNN to act as an emulator for our population model, interpolating across parameters at both a four-dimensional source level and six-dimensional population level. By approximating the detection probabilities and recording the merger generations of individual sources in the simulated populations, we also constructed DNNs to predict, respectively, the fraction of detectable events and the generational branching ratios across the population-level parameter space. These applications of deep learning are then combined with (rate-marginalized) hierarchical Bayesian inference of the events in GWTC-3, performed here with nested sampling, to make measurements of the population-level parameters and reconstruct the astrophysical distribution of merging stellar-mass binary BHs.

This work serves as a showcase of the developments made possible by combining advanced techniques from the fields of deep learning and statistical analysis, applied within the context of GW astrophysics. Our deep-learning population pipeline, which we applied to the case study of simple simulations of hierarchical stellar-mass BH mergers, is thus ready to be used in conjunction with more sophisticated simulated populations. Combined with the state of the art in population synthesis, we will be able to constrain the properties of progenitor formation environments by directly comparing GW data with higher-dimensional models of binary evolution.

ACKNOWLEDGMENTS

We thank Daria Gangardt, Nicola Giacobbo, Nathan Steinle, Chris Moore, Geraint Pratten, Riccardo Buscicchio, Alberto Vecchio, Tom Callister, Mike Zevin, and Kaze Wong for useful discussions. M. M. and D. G. are supported by European Union's H2020 ERC Starting Grant No. 945155-GWmining, Cariplo Foundation Grant No. 2021-0555, Leverhulme Trust Grant No. RPG-2019-350, and Royal Society Grant No. RGS-R2-202004. M. M. acknowledges networking support by the European COST Action CA16104–GWverse. S. R. T. acknowledges support from NSF Grant No. AST-2007993, the NANOGrav Physics Frontier Center NSF Grant No. PHY-2020265, an NSF CAREER Award No. PHY-2146016, and a Vanderbilt University College of Arts & Science Dean's Faculty Fellowship. Computational work was performed on the University of Birmingham BlueBEAR cluster, the Baskerville Tier 2 HPC service funded by EPSRC Grant No. EP/T022221/1, and at CINECA with allocations through INFN, Bicocca, and ISCRA Type-B Project No. HP10BEQ9JB.

APPENDIX: EVENT SELECTION

In Table V, we report all of the GW events that enter our population analysis. We include only the binary BH mergers with FAR < 1 yr^{-1} in at least one of the detection pipelines. The 69 events are equivalent to the 76 listed in Table I of Ref. [31], excluding those that potentially contain a neutron star (GW170817, GW190425_081805, GW190426_152155, GW190814, GW190917_114630, GW200105_162426, and GW200115_042309).

TABLE V. Binary BH mergers that enter our analysis. Each event passes the cut FAR < 1 yr^{-1} in at least one of the searches. We exclude any events that potentially contain a neutron star. For each event, we list the catalog in which it was first reported and the minimum FAR. We list the medians and 90% symmetric intervals for the chirp mass M_c, mass ratio q, effective aligned spin $\chi_{\rm eff}$, and effective precessing spin $\chi_{\rm p}$. The reference frequency used to measure $\chi_{\rm p}$ is 20 Hz for all events except GW190521_030229, which is measured at 11 Hz.

Event	Catalog	min FAR(yr^{-1})	$M_c/M_\odot$	q	$\chi_{\rm eff}$	χ_p
GW150914	GWTC-1	$< 1 \times 10^{-5}$	$28.38^{+1.56}_{-1.47}$	$0.86^{+0.12}_{-0.19}$	$-0.04^{+0.11}_{-0.11}$	$0.34^{+0.50}_{-0.27}$
GW151012	GWTC-1	7.92×10^{-3}	$15.27^{+1.55}_{-1.11}$	$0.71^{+0.26}_{-0.37}$	$0.01^{+0.22}_{-0.17}$	$0.34^{+0.46}_{-0.26}$
GW151226	GWTC-1	$< 1 \times 10^{-5}$	$8.90^{+0.32}_{-0.30}$	$0.66^{+0.30}_{-0.33}$	$0.17^{+0.13}_{-0.06}$	$0.44^{+0.29}_{-0.27}$
GW170104	GWTC-1	$< 1 \times 10^{-5}$	$21.62^{+2.04}_{-1.84}$	$0.70^{+0.26}_{-0.25}$	$-0.05^{+0.16}_{-0.20}$	$0.39^{+0.39}_{-0.30}$
GW170608	GWTC-1	$< 1 \times 10^{-5}$	$7.95^{+0.19}_{-0.18}$	$0.75^{+0.22}_{-0.34}$	$0.04^{+0.13}_{-0.05}$	$0.32^{+0.40}_{-0.25}$
GW170729	GWTC-1	1.80×10^{-1}	$35.23^{+6.33}_{-4.87}$	$0.66^{+0.30}_{-0.28}$	$0.32^{+0.22}_{-0.26}$	$0.44^{+0.46}_{-0.30}$
GW170809	GWTC-1	$< 1 \times 10^{-5}$	$24.93^{+2.19}_{-1.63}$	$0.73^{+0.25}_{-0.25}$	$0.05^{+0.17}_{-0.15}$	$0.37^{+0.41}_{-0.27}$
GW170814	GWTC-1	$< 1 \times 10^{-5}$	$24.17^{+1.45}_{-1.23}$	$0.86^{+0.13}_{-0.22}$	$0.07^{+0.12}_{-0.12}$	$0.50^{+0.50}_{-0.40}$
GW170818	GWTC-1	$< 1 \times 10^{-5}$	$26.67^{+1.99}_{-1.75}$	$0.78^{+0.19}_{-0.24}$	$-0.10^{+0.17}_{-0.21}$	$0.51^{+0.42}_{-0.34}$
GW170823	GWTC-1	$< 1 \times 10^{-5}$	$29.29^{+4.49}_{-3.25}$	$0.77^{+0.20}_{-0.27}$	$0.05^{+0.20}_{-0.21}$	$0.46^{+0.50}_{-0.34}$
GW190408_181802	GWTC-2	$< 1 \times 10^{-5}$	$18.32^{+1.87}_{-1.23}$	$0.75^{+0.21}_{-0.25}$	$-0.03^{+0.13}_{-0.19}$	$0.38^{+0.39}_{-0.30}$
GW190412	GWTC-2	$< 1 \times 10^{-5}$	$13.26^{+0.40}_{-0.33}$	$0.28^{+0.12}_{-0.06}$	$0.25^{+0.08}_{-0.10}$	$0.32^{+0.18}_{-0.16}$
GW190413_052954	GWTC-2	8.17×10^{-1}	$24.70^{+5.43}_{-4.03}$	$0.68^{+0.28}_{-0.26}$	$-0.01^{+0.28}_{-0.35}$	$0.42^{+0.46}_{-0.31}$
GW190413_134308	GWTC-2	1.81×10^{-1}	$33.00^{+8.35}_{-5.29}$	$0.69^{+0.27}_{-0.32}$	$-0.03^{+0.24}_{-0.29}$	$0.56^{+0.48}_{-0.42}$
GW190421_213856	GWTC-2	2.83×10^{-3}	$31.18^{+5.90}_{-4.32}$	$0.79^{+0.18}_{-0.31}$	$-0.05^{+0.22}_{-0.27}$	$0.47^{+0.51}_{-0.35}$
GW190503_185404	GWTC-2	$< 1 \times 10^{-5}$	$30.09^{+4.52}_{-4.23}$	$0.66^{+0.28}_{-0.24}$	$-0.03^{+0.21}_{-0.26}$	$0.40^{+0.44}_{-0.30}$
GW190512_180714	GWTC-2	$< 1 \times 10^{-5}$	$14.60^{+1.27}_{-1.00}$	$0.53^{+0.37}_{-0.17}$	$0.03^{+0.12}_{-0.14}$	$0.22^{+0.35}_{-0.17}$
GW190513_205428	GWTC-2	$< 1 \times 10^{-5}$	$21.57^{+3.78}_{-1.92}$	$0.52^{+0.41}_{-0.19}$	$0.11^{+0.28}_{-0.17}$	$0.31^{+0.38}_{-0.23}$

(Table continued)

TABLE V. *(Continued)*

Event	Catalog	min FAR(yr^{-1})	$M_c/M_\odot$	q	$\chi_{\rm eff}$	χ_p
GW190517_055101	GWTC-2	3.47×10^{-4}	$26.58^{+3.78}_{-3.93}$	$0.68^{+0.27}_{-0.28}$	$0.52^{+0.19}_{-0.20}$	$0.51^{+0.37}_{-0.30}$
GW190519_153544	GWTC-2	$< 1 \times 10^{-5}$	$44.43^{+6.26}_{-7.28}$	$0.61^{+0.26}_{-0.19}$	$0.31^{+0.20}_{-0.22}$	$0.46^{+0.38}_{-0.29}$
GW190521	GWTC-2	$< 1 \times 10^{-5}$	$69.11^{+17.31}_{-10.55}$	$0.75^{+0.22}_{-0.34}$	$0.03^{+0.31}_{-0.39}$	$0.71^{+0.55}_{-0.47}$
GW190521_074359	GWTC-2	1.00×10^{-2}	$32.01^{+3.29}_{-2.44}$	$0.77^{+0.19}_{-0.20}$	$0.09^{+0.10}_{-0.13}$	$0.39^{+0.34}_{-0.28}$
GW190527_092055	GWTC-2	2.28×10^{-1}	$24.30^{+9.13}_{-4.26}$	$0.64^{+0.32}_{-0.32}$	$0.11^{+0.27}_{-0.27}$	$0.45^{+0.50}_{-0.34}$
GW190602_175927	GWTC-2	$< 1 \times 10^{-5}$	$48.97^{+8.88}_{-8.72}$	$0.70^{+0.25}_{-0.33}$	$0.07^{+0.27}_{-0.24}$	$0.42^{+0.53}_{-0.30}$
GW190620_030421	GWTC-2	1.12×10^{-2}	$38.22^{+8.10}_{-6.45}$	$0.62^{+0.33}_{-0.27}$	$0.32^{+0.23}_{-0.24}$	$0.44^{+0.41}_{-0.29}$
GW190630_185205	GWTC-2	$< 1 \times 10^{-5}$	$24.98^{+2.01}_{-2.16}$	$0.68^{+0.27}_{-0.22}$	$0.09^{+0.12}_{-0.12}$	$0.32^{+0.34}_{-0.23}$
GW190701_203306	GWTC-2	5.71×10^{-3}	$40.25^{+5.40}_{-5.12}$	$0.77^{+0.21}_{-0.32}$	$-0.07^{+0.22}_{-0.30}$	$0.41^{+0.51}_{-0.30}$
GW190706_222641	GWTC-2	$< 1 \times 10^{-5}$	$42.81^{+10.20}_{-7.17}$	$0.58^{+0.34}_{-0.25}$	$0.28^{+0.25}_{-0.29}$	$0.40^{+0.45}_{-0.29}$
GW190707_093326	GWTC-2	$< 1 \times 10^{-5}$	$8.47^{+0.63}_{-0.42}$	$0.73^{+0.22}_{-0.24}$	$-0.03^{+0.09}_{-0.08}$	$0.26^{+0.35}_{-0.21}$
GW190708_232457	GWTC-2	3.09×10^{-4}	$13.15^{+0.89}_{-0.66}$	$0.75^{+0.22}_{-0.27}$	$0.02^{+0.09}_{-0.07}$	$0.30^{+0.39}_{-0.24}$
GW190719_215514	GWTC-2	6.31×10^{-1}	$23.47^{+6.62}_{-4.02}$	$0.58^{+0.36}_{-0.30}$	$0.32^{+0.28}_{-0.31}$	$0.43^{+0.40}_{-0.30}$
GW190720_000836	GWTC-2	$< 1 \times 10^{-5}$	$8.79^{+0.59}_{-0.77}$	$0.63^{+0.31}_{-0.28}$	$0.18^{+0.13}_{-0.11}$	$0.30^{+0.35}_{-0.20}$
GW190725_174728	GWTC-2.1	4.58×10^{-1}	$7.44^{+0.56}_{-0.54}$	$0.57^{+0.37}_{-0.31}$	$-0.04^{+0.26}_{-0.14}$	$0.38^{+0.49}_{-0.29}$
GW190727_060333	GWTC-2	$< 1 \times 10^{-5}$	$28.61^{+5.33}_{-3.83}$	$0.79^{+0.18}_{-0.31}$	$0.11^{+0.25}_{-0.25}$	$0.48^{+0.50}_{-0.37}$
GW190728_064510	GWTC-2	$< 1 \times 10^{-5}$	$8.62^{+0.52}_{-0.33}$	$0.69^{+0.27}_{-0.30}$	$0.12^{+0.13}_{-0.06}$	$0.29^{+0.30}_{-0.20}$
GW190731_140936	GWTC-2	3.35×10^{-1}	$29.55^{+6.68}_{-5.19}$	$0.72^{+0.25}_{-0.31}$	$0.05^{+0.25}_{-0.23}$	$0.41^{+0.51}_{-0.31}$
GW190803_022701	GWTC-2	7.32×10^{-2}	$27.26^{+5.48}_{-3.97}$	$0.74^{+0.23}_{-0.31}$	$-0.03^{+0.24}_{-0.27}$	$0.45^{+0.54}_{-0.33}$
GW190805_211137	GWTC-2.1	6.28×10^{-1}	$33.68^{+9.75}_{-7.23}$	$0.68^{+0.27}_{-0.32}$	$0.35^{+0.30}_{-0.36}$	$0.55^{+0.47}_{-0.36}$
GW190828_063405	GWTC-2	$< 1 \times 10^{-5}$	$24.95^{+3.48}_{-2.15}$	$0.82^{+0.15}_{-0.22}$	$0.19^{+0.16}_{-0.16}$	$0.43^{+0.45}_{-0.31}$
GW190828_065509	GWTC-2	$< 1 \times 10^{-5}$	$13.36^{+1.20}_{-0.97}$	$0.43^{+0.38}_{-0.16}$	$0.08^{+0.16}_{-0.16}$	$0.29^{+0.39}_{-0.23}$
GW190910_112807	GWTC-2	2.87×10^{-3}	$34.25^{+4.21}_{-3.96}$	$0.82^{+0.15}_{-0.23}$	$0.02^{+0.18}_{-0.18}$	$0.40^{+0.40}_{-0.31}$
GW190915_235702	GWTC-2	$< 1 \times 10^{-5}$	$25.05^{+3.02}_{-2.59}$	$0.79^{+0.19}_{-0.29}$	$0.02^{+0.19}_{-0.23}$	$0.57^{+0.41}_{-0.41}$
GW190924_021846	GWTC-2	$< 1 \times 10^{-5}$	$5.76^{+0.26}_{-0.21}$	$0.60^{+0.33}_{-0.25}$	$0.02^{+0.14}_{-0.08}$	$0.22^{+0.33}_{-0.17}$
GW190925_232845	GWTC-2.1	7.20×10^{-3}	$15.78^{+1.06}_{-0.95}$	$0.74^{+0.22}_{-0.29}$	$0.11^{+0.16}_{-0.14}$	$0.40^{+0.42}_{-0.29}$
GW190929_012149	GWTC-2	1.55×10^{-1}	$34.09^{+9.34}_{-6.67}$	$0.35^{+0.36}_{-0.17}$	$0.01^{+0.27}_{-0.28}$	$0.38^{+0.43}_{-0.28}$
GW190930_133541	GWTC-2	1.23×10^{-2}	$8.51^{+0.49}_{-0.48}$	$0.66^{+0.28}_{-0.36}$	$0.14^{+0.19}_{-0.13}$	$0.34^{+0.34}_{-0.24}$
GW191103_012549	GWTC-3	4.58×10^{-1}	$8.34^{+0.65}_{-0.57}$	$0.67^{+0.29}_{-0.36}$	$0.21^{+0.16}_{-0.10}$	$0.41^{+0.40}_{-0.26}$
GW191105_143521	GWTC-3	1.18×10^{-2}	$7.81^{+0.61}_{-0.45}$	$0.72^{+0.24}_{-0.30}$	$-0.02^{+0.12}_{-0.10}$	$0.30^{+0.45}_{-0.24}$
GW191109_010717	GWTC-3	1.80×10^{-4}	$47.08^{+9.52}_{-7.33}$	$0.72^{+0.22}_{-0.23}$	$-0.30^{+0.39}_{-0.29}$	$0.60^{+0.67}_{-0.36}$
GW191127_050227	GWTC-3	2.49×10^{-1}	$29.73^{+11.70}_{-9.16}$	$0.47^{+0.46}_{-0.36}$	$0.17^{+0.34}_{-0.36}$	$0.52^{+0.44}_{-0.41}$
GW191129_134029	GWTC-3	$< 1 \times 10^{-5}$	$7.30^{+0.43}_{-0.28}$	$0.64^{+0.30}_{-0.29}$	$0.06^{+0.16}_{-0.07}$	$0.27^{+0.36}_{-0.20}$
GW191204_171526	GWTC-3	$< 1 \times 10^{-5}$	$8.56^{+0.38}_{-0.27}$	$0.69^{+0.26}_{-0.26}$	$0.16^{+0.08}_{-0.05}$	$0.40^{+0.35}_{-0.26}$
GW191215_223052	GWTC-3	$< 1 \times 10^{-5}$	$18.34^{+2.20}_{-1.65}$	$0.73^{+0.24}_{-0.27}$	$-0.04^{+0.17}_{-0.21}$	$0.52^{+0.43}_{-0.38}$
GW191216_213338	GWTC-3	$< 1 \times 10^{-5}$	$8.33^{+0.22}_{-0.19}$	$0.62^{+0.32}_{-0.28}$	$0.11^{+0.13}_{-0.06}$	$0.24^{+0.33}_{-0.16}$
GW191222_033537	GWTC-3	$< 1 \times 10^{-5}$	$33.82^{+7.04}_{-5.01}$	$0.80^{+0.18}_{-0.32}$	$-0.04^{+0.19}_{-0.25}$	$0.41^{+0.49}_{-0.30}$
GW191230_180458	GWTC-3	5.02×10^{-2}	$36.37^{+8.13}_{-5.55}$	$0.76^{+0.22}_{-0.33}$	$-0.06^{+0.27}_{-0.30}$	$0.53^{+0.51}_{-0.39}$
GW200112_155838	GWTC-3	$< 1 \times 10^{-5}$	$27.34^{+2.59}_{-2.01}$	$0.80^{+0.17}_{-0.26}$	$0.06^{+0.15}_{-0.14}$	$0.39^{+0.39}_{-0.30}$
GW200128_022011	GWTC-3	4.29×10^{-3}	$32.04^{+7.55}_{-5.53}$	$0.80^{+0.17}_{-0.31}$	$0.12^{+0.24}_{-0.25}$	$0.60^{+0.54}_{-0.42}$
GW200129_065458	GWTC-3	$< 1 \times 10^{-5}$	$27.16^{+2.07}_{-2.28}$	$0.85^{+0.12}_{-0.41}$	$0.11^{+0.11}_{-0.16}$	$0.50^{+0.47}_{-0.35}$

(Table continued)

TABLE V. *(Continued)*

Event	Catalog	min FAR(yr^{-1})	$M_c/M_\odot$	q	χ_{eff}	χ_p
GW200202_154313	GWTC-3	$< 1 \times 10^{-5}$	$7.49^{+0.23}_{-0.20}$	$0.72^{+0.25}_{-0.32}$	$0.04^{+0.14}_{-0.06}$	$0.29^{+0.41}_{-0.22}$
GW200208_130117	GWTC-3	3.11×10^{-4}	$27.71^{+3.55}_{-3.05}$	$0.73^{+0.24}_{-0.28}$	$-0.07^{+0.21}_{-0.26}$	$0.38^{+0.46}_{-0.29}$
GW200209_085452	GWTC-3	4.64×10^{-2}	$26.77^{+5.93}_{-4.17}$	$0.79^{+0.19}_{-0.32}$	$-0.12^{+0.24}_{-0.30}$	$0.52^{+0.57}_{-0.39}$
GW200216_220804	GWTC-3	3.50×10^{-1}	$32.93^{+9.15}_{-8.51}$	$0.61^{+0.34}_{-0.40}$	$0.10^{+0.34}_{-0.38}$	$0.46^{+0.48}_{-0.35}$
GW200219_094415	GWTC-3	9.94×10^{-4}	$27.72^{+5.74}_{-3.93}$	$0.77^{+0.20}_{-0.32}$	$-0.08^{+0.23}_{-0.29}$	$0.47^{+0.50}_{-0.34}$
GW200224_222234	GWTC-3	$< 1 \times 10^{-5}$	$31.09^{+3.18}_{-2.56}$	$0.82^{+0.16}_{-0.26}$	$0.10^{+0.14}_{-0.15}$	$0.49^{+0.46}_{-0.35}$
GW200225_060421	GWTC-3	$< 1 \times 10^{-5}$	$14.21^{+1.47}_{-1.33}$	$0.73^{+0.24}_{-0.28}$	$-0.11^{+0.18}_{-0.28}$	$0.55^{+0.41}_{-0.40}$
GW200302_015811	GWTC-3	1.12×10^{-1}	$23.36^{+4.67}_{-2.92}$	$0.53^{+0.36}_{-0.20}$	$0.01^{+0.25}_{-0.26}$	$0.38^{+0.44}_{-0.29}$
GW200311_115853	GWTC-3	$< 1 \times 10^{-5}$	$26.58^{+2.35}_{-1.92}$	$0.81^{+0.16}_{-0.27}$	$-0.02^{+0.16}_{-0.19}$	$0.44^{+0.45}_{-0.33}$
GW200316_215756	GWTC-3	$< 1 \times 10^{-5}$	$8.75^{+0.65}_{-0.55}$	$0.59^{+0.34}_{-0.38}$	$0.13^{+0.28}_{-0.10}$	$0.31^{+0.38}_{-0.20}$

[1] J. Aasi *et al.* (LIGO Collaboration), Classical Quantum Gravity **32**, 115012 (2015).

[2] F. Acernese *et al.* (Virgo Collaboration), Classical Quantum Gravity **32**, 024001 (2015).

[3] B. P. Abbott *et al.* (LIGO and Virgo Collaborations), Phys. Rev. X **9**, 031040 (2019).

[4] R. Abbott *et al.* (LIGO and Virgo Collaborations), Phys. Rev. X **11**, 021053 (2021).

[5] R. Abbott *et al.* (LIGO and Virgo Collaborations), arXiv: 2108.01045.

[6] R. Abbott *et al.* (LIGO, Virgo, and KAGRA Collaborations), arXiv:2111.03606.

[7] A. H. Nitz, C. D. Capano, S. Kumar, Y.-F. Wang, S. Kastha, M. Schäfer, R. Dhurkunde, and M. Cabero, Astrophys. J. **922**, 76 (2021).

[8] A. H. Nitz, S. Kumar, Y.-F. Wang, S. Kastha, S. Wu, M. Schäfer, R. Dhurkunde, and C. D. Capano, arXiv:2112.06878.

[9] S. Olsen, T. Venumadhav, J. Mushkin, J. Roulet, B. Zackay, and M. Zaldarriaga, Phys. Rev. D **106**, 043009 (2022).

[10] K. A. Postnov and L. R. Yungelson, Living Rev. Relativity **17**, 3 (2014).

[11] M. J. Benacquista and J. M. B. Downing, Living Rev. Relativity **16**, 4 (2013).

[12] R. Farmer, M. Renzo, S. E. de Mink, P. Marchant, and S. Justham, Astrophys. J. **887**, 53 (2019).

[13] S. E. Woosley and A. Heger, Astrophys. J. Lett. **912**, L31 (2021).

[14] V. Kalogera, Astrophys. J. **541**, 319 (2000).

[15] D. Gerosa, E. Berti, R. O'Shaughnessy, K. Belczynski, M. Kesden, D. Wysocki, and W. Gladysz, Phys. Rev. D **98**, 084036 (2018).

[16] N. Steinle and M. Kesden, Phys. Rev. D **103**, 063032 (2021).

[17] D. Gerosa and M. Fishbach, Nat. Astron. **5**, 749 (2021).

[18] T. Bogdanović, C. S. Reynolds, and M. C. Miller, Astrophys. J. Lett. **661**, L147 (2007).

[19] D. Gerosa, M. Kesden, U. Sperhake, E. Berti, and R. O'Shaughnessy, Phys. Rev. D **92**, 064016 (2015).

[20] I. Mandel and A. Farmer, Phys. Rep. **955**, 1 (2022).

[21] I. Mandel and F. S. Broekgaarden, Living Rev. Relativity **25**, 1 (2022).

[22] M. Mapelli, in *Handbook of Gravitational Wave Astronomy* (Springer, New York, 2021), p. 4.

[23] J. W. Barrett, S. M. Gaebel, C. J. Neijssel, A. Vigna-Gómez, S. Stevenson, C. P. L. Berry, W. M. Farr, and I. Mandel, Mon. Not. R. Astron. Soc. **477**, 4685 (2018).

[24] K. Belczynski, A. Romagnolo, A. Olejak, J. Klencki, D. Chattopadhyay, S. Stevenson, M. Coleman Miller, J. P. Lasota, and P. A. Crowther, Astrophys. J. **925**, 69 (2022).

[25] I. Mandel, W. M. Farr, and J. R. Gair, Mon. Not. R. Astron. Soc. **486**, 1086 (2019).

[26] S. Vitale, D. Gerosa, W. M. Farr, and S. R. Taylor, in *Handbook of Gravitational Wave Astronomy* (Springer, New York, 2022), p. 45.

[27] M. Zevin, S. S. Bavera, C. P. L. Berry, V. Kalogera, T. Fragos, P. Marchant, C. L. Rodriguez, F. Antonini, D. E. Holz, and C. Pankow, Astrophys. J. **910**, 152 (2021).

[28] Y. Bouffanais, M. Mapelli, F. Santoliquido, N. Giacobbo, U. N. Di Carlo, S. Rastello, M. C. Artale, and G. Iorio, Mon. Not. R. Astron. Soc. **507**, 5224 (2021).

[29] B. P. Abbott *et al.* (LIGO and Virgo Collaborations), Astrophys. J. Lett. **882**, L24 (2019).

[30] R. Abbott *et al.* (LIGO and Virgo Collaborations), Astrophys. J. Lett. **913**, L7 (2021).

[31] R. Abbott *et al.* (LIGO, Virgo, and KAGRA Collaborations), arXiv:2111.03634.

[32] T. A. Callister, C.-J. Haster, K. K. Y. Ng, S. Vitale, and W. M. Farr, Astrophys. J. Lett. **922**, L5 (2021).

[33] G. Franciolini and P. Pani, Phys. Rev. D **105**, 123024 (2022).

[34] J. Roulet, H. S. Chia, S. Olsen, L. Dai, T. Venumadhav, B. Zackay, and M. Zaldarriaga, Phys. Rev. D **104**, 083010 (2021).

[35] S. Galaudage, C. Talbot, T. Nagar, D. Jain, E. Thrane, and I. Mandel, Astrophys. J. Lett. **921**, L15 (2021).

[36] M. Mould and D. Gerosa, Phys. Rev. D **105**, 024076 (2022).

[37] I. M. Romero-Shaw, E. Thrane, and P. D. Lasky, Pub. Astron. Soc. Aust. **39**, e025 (2022).

[38] B. Edelman, Z. Doctor, J. Godfrey, and B. Farr, Astrophys. J. **924**, 101 (2022).

[39] V. Tiwari, Classical Quantum Gravity **38**, 155007 (2021).

[40] I. Mandel, W. M. Farr, A. Colonna, S. Stevenson, P. Tiňo, and J. Veitch, Mon. Not. R. Astron. Soc. **465**, 3254 (2017).

[41] J. Sadiq, T. Dent, and D. Wysocki, Phys. Rev. D **105**, 123014 (2022).

[42] S. Rinaldi and W. Del Pozzo, Mon. Not. R. Astron. Soc. **509**, 5454 (2021).

[43] S. R. Taylor and D. Gerosa, Phys. Rev. D **98**, 083017 (2018).

[44] K. W. K. Wong and D. Gerosa, Phys. Rev. D **100**, 083015 (2019).

[45] D. H. T. Cheung, K. W. K. Wong, O. A. Hannuksela, T. G. F. Li, and S. Ho, Phys. Rev D **106**, 083014 (2022).

[46] K. W. K. Wong, G. Contardo, and S. Ho, Phys. Rev. D **101**, 123005 (2020).

[47] G. Papamakarios, T. Pavlakou, and I. Murray, arXiv:1705.07057.

[48] K. W. K. Wong, G. Franciolini, V. De Luca, V. Baibhav, E. Berti, P. Pani, and A. Riotto, Phys. Rev. D **103**, 023026 (2021).

[49] K. W. K. Wong, K. Breivik, K. Kremer, and T. Callister, Phys. Rev. D **103**, 083021 (2021).

[50] D. Gerosa, G. Pratten, and A. Vecchio, Phys. Rev. D **102**, 103020 (2020).

[51] C. Talbot and E. Thrane, Astrophys. J. **927**, 76 (2022).

[52] K. W. K. Wong, K. K. Y. Ng, and E. Berti, arXiv:2007.10350.

[53] R. Abbott *et al.* (LIGO and Virgo Collaborations), Phys. Rev. D **102**, 043015 (2020).

[54] D. Gerosa, S. Vitale, and E. Berti, Phys. Rev. Lett. **125**, 101103 (2020).

[55] C. L. Rodriguez, K. Kremer, M. Y. Grudić, Z. Hafen, S. Chatterjee, G. Fragione, A. Lamberts, M. A. S. Martinez, F. A. Rasio, N. Weatherford, and C. S. Ye, Astrophys. J. Lett. **896**, L10 (2020).

[56] A. S. Hamers and M. Safarzadeh, Astrophys. J. **898**, 99 (2020).

[57] M. Safarzadeh and K. Hotokezaka, Astrophys. J. Lett. **897**, L7 (2020).

[58] R. Abbott *et al.* (LIGO and Virgo Collaborations), Phys. Rev. Lett. **125**, 101102 (2020).

[59] R. Abbott *et al.* (LIGO and Virgo Collaborations), Astrophys. J. Lett. **900**, L13 (2020).

[60] I. Romero-Shaw, P. D. Lasky, E. Thrane, and J. Calderón Bustillo, Astrophys. J. Lett. **903**, L5 (2020).

[61] C. Kimball, C. Talbot, C. P. L. Berry, M. Zevin, E. Thrane, V. Kalogera, R. Buscicchio, M. Carney, T. Dent, H. Middleton, E. Payne, J. Veitch, and D. Williams, Astrophys. J. Lett. **915**, L35 (2021).

[62] G. Fragione, A. Loeb, and F. A. Rasio, Astrophys. J. Lett. **902**, L26 (2020).

[63] M. Mapelli, M. Dall'Amico, Y. Bouffanais, N. Giacobbo, M. Arca Sedda, M. C. Artale, A. Ballone, U. N. Di Carlo, G. Iorio, F. Santoliquido, and S. Torniamenti, Mon. Not. R. Astron. Soc. **505**, 339 (2021).

[64] J. Samsing, I. Bartos, D. J. D'Orazio, Z. Haiman, B. Kocsis, N. W. C. Leigh, B. Liu, M. E. Pessah, and H. Tagawa, arXiv:2010.09765.

[65] M. Arca-Sedda, F. P. Rizzuto, T. Naab, J. Ostriker, M. Giersz, and R. Spurzem, Astrophys. J. **920**, 128 (2021).

[66] M. Fishbach and D. E. Holz, Astrophys. J. Lett. **904**, L26 (2020).

[67] R. Essick, A. Farah, S. Galaudage, C. Talbot, M. Fishbach, E. Thrane, and D. E. Holz, Astrophys. J. **926**, 34 (2022).

[68] R. Abbott *et al.* (LIGO and Virgo Collaborations), Astrophys. J. Lett. **896**, L44 (2020).

[69] B. Liu and D. Lai, Mon. Not. R. Astron. Soc. **502**, 2049 (2021).

[70] H. Tagawa, B. Kocsis, Z. Haiman, I. Bartos, K. Omukai, and J. Samsing, Astrophys. J. **908**, 194 (2021).

[71] W. Lu, P. Beniamini, and C. Bonnerot, Mon. Not. R. Astron. Soc. **500**, 1817 (2020).

[72] J. A. González, M. Hannam, U. Sperhake, B. Brügmann, and S. Husa, Phys. Rev. Lett. **98**, 231101 (2007).

[73] M. Campanelli, C. Lousto, Y. Zlochower, and D. Merritt, Astrophys. J. Lett. **659**, L5 (2007).

[74] D. Gerosa, F. Hébert, and L. C. Stein, Phys. Rev. D **97**, 104049 (2018).

[75] D. Gerosa and E. Berti, Phys. Rev. D **100**, 041301 (2019).

[76] V. Baibhav, E. Berti, D. Gerosa, M. Mould, and K. W. K. Wong, Phys. Rev. D **104**, 084002 (2021).

[77] D. Gerosa, N. Giacobbo, and A. Vecchio, Astrophys. J. **915**, 56 (2021).

[78] M. D. McKay, R. J. Beckman, and W. J. Conover, Technometrics **21**, 239 (1979).

[79] A. Heger, C. L. Fryer, S. E. Woosley, N. Langer, and D. H. Hartmann, Astrophys. J. **591**, 288 (2003).

[80] S. E. Woosley, S. Blinnikov, and A. Heger, Nature (London) **450**, 390 (2007).

[81] K. Belczynski, A. Heger, W. Gladysz, A. J. Ruiter, S. Woosley, G. Wiktorowicz, H. Y. Chen, T. Bulik, R. O'Shaughnessy, D. E. Holz, C. L. Fryer, and E. Berti, Astron. Astrophys. **594**, A97 (2016).

[82] R. Farmer, M. Renzo, S. E. de Mink, M. Fishbach, and S. Justham, Astrophys. J. Lett. **902**, L36 (2020).

[83] P. Marchant and T. J. Moriya, Astron. Astrophys. **640**, L18 (2020).

[84] J. R. Rice and B. Zhang, Astrophys. J. **908**, 59 (2021).

[85] M. Safarzadeh and Z. Haiman, Astrophys. J. Lett. **903**, L21 (2020).

[86] Z. Roupas and D. Kazanas, Astron. Astrophys. **632**, L8 (2019).

[87] P. Natarajan, Mon. Not. R. Astron. Soc. **501**, 1413 (2020).

[88] L. A. C. van Son, S. E. De Mink, F. S. Broekgaarden, M. Renzo, S. Justham, E. Laplace, J. Morán-Fraile, D. D. Hendriks, and R. Farmer, Astrophys. J. **897**, 100 (2020).

[89] K. Belczynski, R. Hirschi, E. A. Kaiser, J. Liu, J. Casares, Y. Lu, R. O'Shaughnessy, A. Heger, S. Justham, and R. Soria, Astrophys. J. **890**, 113 (2020).

[90] J. S. Vink, E. R. Higgins, A. A. C. Sander, and G. N. Sabhahit, Mon. Not. R. Astron. Soc. **504**, 146 (2021).

[91] A. Tanikawa, H. Susa, T. Yoshida, A. A. Trani, and T. Kinugawa, Astrophys. J. **910**, 30 (2021).

[92] E. Farrell, J. H. Groh, R. Hirschi, L. Murphy, E. Kaiser, S. Ekström, C. Georgy, and G. Meynet, Mon. Not. R. Astron. Soc. **502**, L40 (2021).

[93] T. Kinugawa, T. Nakamura, and H. Nakano, Mon. Not. R. Astron. Soc. **501**, L49 (2020).

[94] G. Costa, A. Bressan, M. Mapelli, P. Marigo, G. Iorio, and M. Spera, Mon. Not. R. Astron. Soc. **501**, 4514 (2021).

[95] J. Fuller and L. Ma, Astrophys. J. Lett. **881**, L1 (2019).

[96] K. Belczynski *et al.*, Astron. Astrophys. **636**, A104 (2020).

[97] O. Y. Gnedin, H. Zhao, J. E. Pringle, S. M. Fall, M. Livio, and G. Meylan, Astrophys. J. Lett. **568**, L23 (2002).

[98] D. Merritt, M. Milosavljević, M. Favata, S. A. Hughes, and D. E. Holz, Astrophys. J. Lett. **607**, L9 (2004).

[99] F. Antonini and F. A. Rasio, Astrophys. J. **831**, 187 (2016).

[100] D. Gerosa and M. Kesden, Phys. Rev. D **93**, 124066 (2016).

[101] E. Barausse, V. Morozova, and L. Rezzolla, Astrophys. J. **758**, 63 (2012).

[102] E. Barausse and L. Rezzolla, Astrophys. J. Lett. **704**, L40 (2009).

[103] F. Hofmann, E. Barausse, and L. Rezzolla, Astrophys. J. Lett. **825**, L19 (2016).

[104] C. O. Lousto and Y. Zlochower, Phys. Rev. D **77**, 044028 (2008).

[105] C. O. Lousto, Y. Zlochower, M. Dotti, and M. Volonteri, Phys. Rev. D **85**, 084015 (2012).

[106] C. O. Lousto and Y. Zlochower, Phys. Rev. D **87**, 084027 (2013).

[107] J. D. Bekenstein, Astrophys. J. **183**, 657 (1973).

[108] M. J. Fitchett, Mon. Not. R. Astron. Soc. **203**, 1049 (1983).

[109] É. Racine, Phys. Rev. D **78**, 044021 (2008).

[110] D. Gerosa, M. Mould, D. Gangardt, P. Schmidt, G. Pratten, and L. M. Thomas, Phys. Rev. D **103**, 064067 (2021).

[111] D. Gerosa and E. Berti, Phys. Rev. D **95**, 124046 (2017).

[112] F. Pretorius, Phys. Rev. Lett. **95**, 121101 (2005).

[113] B. P. Abbott *et al.* (LIGO and Virgo Collaborations), Phys. Rev. X **6**, 041015 (2016).

[114] M. Fishbach, D. E. Holz, and W. M. Farr, Astrophys. J. Lett. **863**, L41 (2018).

[115] P. Virtanen *et al.*, Nat. Methods **17**, 261 (2020).

[116] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, London, 1986).

[117] J. S. Speagle, Mon. Not. R. Astron. Soc. **493**, 3132 (2020).

[118] G. Ashton *et al.*, Astrophys. J. Suppl. Ser. **241**, 27 (2019).

[119] I. M. Romero-Shaw, C. Talbot, S. Biscoveanu, V. D'Emilio *et al.*, Mon. Not. R. Astron. Soc. **499**, 3295 (2020).

[120] T. Odland, KDEpy 10.5281/zenodo.2392268 (2018).

[121] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, Ann. Stat. **38**, 2916 (2010).

[122] S. J. Sheather and M. C. Jones, J. R. Stat. Soc. Ser. B Stat. Methodol. **53**, 683 (1991), http://www.jstor.org/stable/2345597.

[123] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (USENIX Association, Savannah, GA, 2016), pp. 265–283.

[124] B. Xu, N. Wang, T. Chen, and M. Li, arXiv:1505.00853.

[125] D. P. Kingma and J. Ba, in *3rd International Conference for Learning Representations* (Conference Track Proceedings, San Diego, CA, 2014), arXiv:1412.6980.

[126] E. Hellinger, J. Reine Angew. Math. **1909**, 210 (1909).

[127] C. J. Moore and D. Gerosa, Phys. Rev. D **104**, 083008 (2021).

[128] L. S. Finn and D. F. Chernoff, Phys. Rev. D **47**, 2198 (1993).

[129] L. S. Finn, Phys. Rev. D **53**, 2878 (1996).

[130] D. Gerosa, gwdet 10.5281/zenodo.889966 (2017).

[131] A. Nitz, I. Harry, D. Brown, C. M. Biwer *et al.*, pycbc 10.5281/zenodo.4556907 (2021).

[132] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Phys. Rev. Lett. **113**, 151101 (2014).

[133] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. J. Forteza, and A. Bohé, Phys. Rev. D **93**, 044006 (2016).

[134] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. J. Forteza, and A. Bohé, Phys. Rev. D **93**, 044007 (2016).

[135] B. P. Abbott *et al.* (LIGO and Virgo Collaborations), Living Rev. Relativity **21**, 3 (2018).

[136] B. P. Abbott *et al.* (LIGO and Virgo Collaborations), Astrophys. J. Suppl. Ser. **227**, 14 (2016).

[137] A. Sesana, J. Gair, E. Berti, and M. Volonteri, Phys. Rev. D **83**, 044036 (2011).

[138] Y. Bouffanais, M. Mapelli, D. Gerosa, U. N. Di Carlo, N. Giacobbo, E. Berti, and V. Baibhav, Astrophys. J. **886**, 25 (2019).

[139] A. Toubiana, K. W. K. Wong, S. Babak, E. Barausse, E. Berti, J. R. Gair, S. Marsat, and S. R. Taylor, Phys. Rev. D **104**, 083027 (2021).

[140] P. Kroupa, Mon. Not. R. Astron. Soc. **322**, 231 (2001).

[141] K. Belczynski, Astrophys. J. Lett. **905**, L15 (2020).

[142] V. Tiwari and S. Fairhurst, Astrophys. J. Lett. **913**, L19 (2021).

[143] M. Zaldarriaga, D. Kushnir, and J. A. Kollmeier, Mon. Not. R. Astron. Soc. **473**, 4174 (2018).

[144] S. S. Bavera, T. Fragos, Y. Qin, E. Zapartas, C. J. Neijssel, I. Mandel, A. Batta, S. M. Gaebel, C. Kimball, and S. Stevenson, Astron. Astrophys. **635**, A97 (2020).

[145] P. Schmidt, F. Ohme, and M. Hannam, Phys. Rev. D **91**, 024043 (2015).

[146] C. Henshaw, R. O'Shaughnessy, and L. Cadonati, Classical Quantum Gravity **39**, 125003 (2022).

[147] Z. Doctor, D. Wysocki, R. O'Shaughnessy, D. E. Holz, and B. Farr, Astrophys. J. **893**, 35 (2020).

[148] C. Kimball, C. Talbot, C. P. L. Berry, M. Carney, M. Zevin, E. Thrane, and V. Kalogera, Astrophys. J. **900**, 177 (2020).

[149] B. McKernan, K. E. S. Ford, R. O'Shaugnessy, and D. Wysocki, Mon. Not. R. Astron. Soc. **494**, 1203 (2020).

[150] M. Fishbach, C. Kimball, and V. Kalogera, Astrophys. J. Lett. **935**, L26 (2022).

[151] A. van den Oord *et al.*, arXiv:1711.10433.

[152] L. Valentin Jospin, W. Buntine, F. Boussaid, H. Laga, and M. Bennamoun, arXiv:2007.06823.