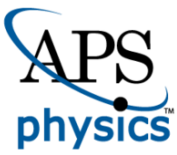

Mining gravitational-wave catalogs to understand binary stellar evolution: A new hierarchical Bayesian framework

Stephen R. Taylor and Davide Gerosa

Phys. Rev. D **98**, 083017 — Published 18 October 2018

DOI: 10.1103/PhysRevD.98.083017

Mining Gravitational-wave Catalogs To Understand Binary Stellar Evolution: A New Hierarchical Bayesian Framework

Stephen R. Taylor[1, 2, *] and Davide Gerosa[1, †]

[1] *TAPIR 350-17, California Institute of Technology, 1200 E California Boulevard, Pasadena, CA 91125, USA*
[2] *Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA*

(Dated: September 20, 2018)

Catalogs of stellar-mass compact binary systems detected by ground-based gravitational-wave instruments (such as Advanced LIGO and Advanced Virgo) will offer insights into the demographics of progenitor systems and the physics guiding stellar evolution. Existing techniques approach this through phenomenological modeling, discrete model selection, or model mixtures. Instead, we explore a novel technique that mines gravitational-wave catalogs to directly infer posterior probability distributions of the hyper-parameters describing formation and evolutionary scenarios (e.g. progenitor metallicity, kick parameters, and common-envelope efficiency). We use a bank of compact-binary population synthesis simulations to train a Gaussian-process emulator that acts as a prior on observed parameter distributions (e.g. chirp mass, redshift, rate). This emulator slots into a hierarchical population inference framework to extract the underlying astrophysical origins of systems detected by Advanced LIGO and Advanced Virgo. Our method is fast, easily expanded with additional simulations, and can be adapted for training on arbitrary population synthesis codes, as well as different detectors like LISA.

Keywords: gravitational waves, gaussian processes, population synthesis, black holes, data analysis, hierarchical Bayesian modeling, stellar evolution

I. INTRODUCTION

Over the last few years, the Advanced LIGO and Advanced Virgo interferometers have detected gravitational-waves (GWs) emitted during the final inspiral and merger of binary black holes and neutron stars. Among the many fruits of these ongoing searches have been the first direct detection of GWs from binary black-hole (BH) systems [1]; a growing catalog of BHs at various masses, distances, and component spin orientations [2–6]; and the first double neutron-star (NS) merger signal [7], with a plethora of associated multi-messenger electromagnetic follow-up analysis [8]. The expected detection rate of binary BHs and NSs could be tens per year with current detectors [2], and promise a data explosion for future third-generation ground-based interferometers [9]. As we move from the dawn of GW astronomy into its source-rich golden-age, we will be able to perform detailed reconstructions of the demographics of stellar populations, the formation history of compact binary systems, and the physical processes guiding stellar evolution.

There are undoubtedly individual GW detections that can provide invaluable physical and astrophysical insight. For instance, the detection of GW150914 proved that GWs could be directly detected [1] and that GW emission was consistent with GR [10, 11]. Perhaps even more crucially from an astrophysical standpoint, it gave the first irrefutable proof that BHs indeed form binary systems able to merger within a Hubble time. Likewise, the detection and electromagnetic follow-up of GW170817 showed that NS mergers could explain the origin of short gamma-ray bursts [8]; gave insight into the equation of state of nuclear matter [12, 13]; constrained the speed of the graviton to less than one part in 10^{-15} [14]; and even permitted a measurement of the Hubble constant [15]. There will continue to be such "golden" systems offering unique physical insights. For instance, detections with particularly favorable orientations in the future might show signs of spin precession [16].

But even with the small number of GW detections so far, emphasis is already shifting to answering questions about the population properties of GW sources. As we move towards the large-statistics regime of GW astronomy, focus will shift from inferring *parameters* of single sources (masses, spins, redshifts) to characterizing *hyper-parameters* describing formation and evolutionary processes of BH and NS populations.

There are many challenges to understanding the formation channels of GW-detected compact binary systems [17]. Binary stellar evolutionary codes (e.g. [18–25]) have become very detailed, but still suffer from large theoretical uncertainties. To name a few, these include (*i*) the dependence of remnant compact object masses (and thus NS or BH identities) on stellar winds and metallicity; (*ii*) the magnitude of kicks received by BHs and NSs at formation; and (*iii*) the efficiency with which orbital energy can be transferred to a common envelope, thereby tightening a binary. Adding to these uncertainties in classical isolated binary evolution are details of other proposed scenarios involving dynamical interactions with

* NANOGrav Senior Postdoctoral Fellow; srtaylor@caltech.edu
† Einstein Fellow; dgerosa@caltech.edu

other bodies [26]. There is thus much poorly known stellar astrophysics that catalogs of GW detections can be mined for.

Several techniques have been developed to perform GW population inference, ranging from phenomenological parametrized modeling to discrete model selection, with mixture modeling as a blending of the former two. In phenomenological models, the distribution of component masses, spins, and redshifts are reconstructed through relatively simple parametrizations (e.g. [27–33]). Any inference with these models will only be a broad sketch of the complicated process of compact binary formation. Detailed stellar population modeling allows binary stars to be tracked from known astrophysical assumptions all the way through to compact binary formation (or not, depending on conditions). But these are computationally expensive (making real-time simulation runs during Bayesian analysis unfeasible), and are typically performed in small batches for comparisons to observations. This approach has been very successful, showing e.g. that GW150914's stellar progenitor had a metallicity of $\sim 5\% \, Z_\odot$ [34–36]. More systematic approaches have also been taken, where Bayesian model selection is performed on grids of discrete population synthesis simulations, or where simulations are mixed together with weightings inferred from the data [29, 37–40]. Finally, non-parametric methods have been developed to allow recovery of binary parameter distributions that is more agnostic than the parametrized-model approach [41]. These methods recover the bin heights of parameter distribution histograms, typically with Gaussian Process (GP) priors linking the bins to enforce smoothness.

In this paper we present a qualitatively new approach that fuses non-parametric modeling with population-synthesis simulations. In brief, we model histograms of GW parameter distributions with bin heights constrained by informative parametrized-priors built out of population synthesis simulations. This allows us to fully exploit catalogs of GW detections to directly infer the properties of progenitors and the evolutionary path undertaken. Our methods give predictions of rates and parameter distributions of compact-binary systems by interpolating between a set of population-synthesis simulations informed by the data. Crucially, the framework developed here remains agnostic of the specific population synthesis code to used.

We follow a multi-stage process (illustrated in Fig. 1), beginning with a design for the program of simulations across hyper-parameter space, compressing distributions of binary parameters to distill the most important features, and training a GP model to interpolate between the simulation hyper-parameter coordinates. These models are then fed to a hierarchical Bayesian pipeline to recover the joint posterior probability distribution of population hyper-parameters, while incorporating measurement uncertainties in each binary's parameters. GP emulation of computationally-expensive simulations has been used in cosmological matter power spectrum analysis [42, 43], pulsar-timing array GW constraints on supermassive binary BH dynamical environments [44, 45], and has been suggested in principle for stellar-mass binary BH population inference [46]. Here we fully develop this emulation approach, embedding it in a complete end-to-end statistical framework, starting from the simulation program design and following through to GW catalog analysis.

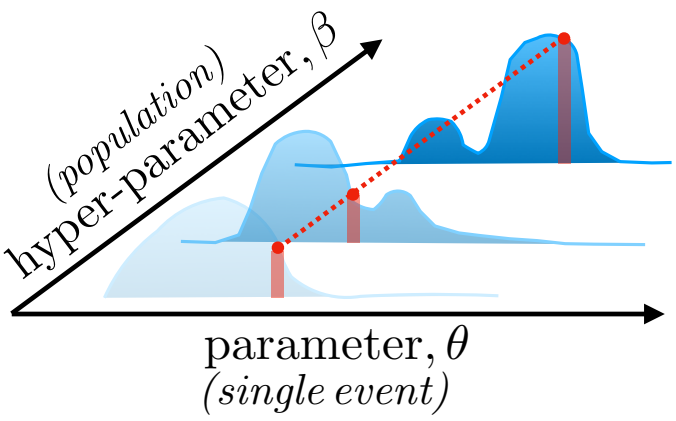

FIG. 1. A schematic representation of interpolating over parameter distributions(θ, e.g. masses, spins, redshift) as a function of population hyper-parameters (β, e.g. progenitor metallicity, common-envelope hardening efficiency, natal kicks, etc.). We carry out a restricted number of population synthesis simulations with different hyper-parameters, where each simulation produces compact binaries distributed over parameter space. These parameter distributions form the training data for our interpolant model. For each bin, pixel, or feature in the parameter distribution, we train a GP interpolant over the hyper-parameter space, allowing us to predict the distribution at any other hyper-parameter coordinate.

This paper is laid out as follows. In Sec. II we describe how to choose locations in the hyper-parameter space where we should perform simulations, how to compress distributions of simulated binary parameters, and how we interpolate over these compressed distributions using GPs. We introduce our inference tools in Sec. III, including Bayesian GW parameter estimation, a scheme to convolve the intrinsic simulated binary distributions with detector selection effects, and a pipeline to perform hierarchical Bayesian inference on catalogs of GW detections. We show our results in Sec. IV, where our entire framework is tested on three case studies that successively increase in complexity and astrophysical realism. These include (i) a toy analytic model, (ii) an example with publicly-available population synthesis simulations, and (iii) finally an example with our custom program of simulations. We provide our conclusions and a discussion of future prospects in Sec. V.

II. STATISTICAL FRAMEWORK

In this Section we describe a statistical framework for choosing points in hyper-parameter space at which to generate simulated astrophysical populations (Sec. II A), defining a data-driven basis for the distributions of population parameters (Sec. II B), and training an interpolation scheme to emulate these parameter distributions (Sec. II C). Our framework closely follows the steps outlined for cosmological matter power spectrum studies in Refs. [42, 43].

A. Simulation design

We need a careful strategy for determining the locations in hyper-parameter space at which to perform the simulations that will eventually be used to train our emulator. While the temptation is to choose an N-dimensional grid-design, this turns out to be highly sub-optimal. The hyper-parameter space dictating stellar-mass binary evolution is $\mathcal{O}(10)$ dimensions, and grid-based designs quickly explode in the number of required simulations. For example, if we choose a simple grid with 3 nodes along each dimension, then in 2-dimensions this is a reasonable choice, requiring 9 simulations in total. However, expanding this to 10 dimensions requires $3^{10} \sim 6 \times 10^4$ simulations, which is a computationally prohibitive step for current population-synthesis codes. The entire purpose of constructing an emulator is to avoid the need for high numbers of costly simulation runs. Furthermore, grid-based designs are poor at covering low-dimensional projections of the full hyper-parameter space. If the distribution of BH masses and spins is dominated by only three hyper-parameters (say progenitor metallicity, natal kicks, and common-envelope efficiency) out of the full 10 dimensional space, then our above-mentioned grid-based design only assigns $3^3 = 27$ unique simulated combinations of these important hyper-parameters out of the total $\sim 6 \times 10^4$ simulations. The opposite case is a purely random design, which however suffers from large regions of sparsely populated hyper-parameter space because random sampling maintains no record of where previous points have been placed.

One thus needs a simulation design that gives good coverage over all lower-dimensional projections of the hyper-parameter space, while simultaneously being sparse enough in the full space to make the program of simulations computationally tractable. A popular solution is given by stratified sampling. If M points are to be drawn, the hyper-parameter volume is first divided into M equally-probable sub-strata, within which random sampling for each point is employed. Specifically, we use space-filling Latin hypercube designs [47], where each sample is the only one permitted to occupy the axis-aligned hyperplane containing it. One must define how many samples are to be drawn at the outset of sampling, and the sampler keeps a record of the position of each past draw. A variant on this technique for integers in the range $[0, 9]$ produces the popular puzzle Sudoku.

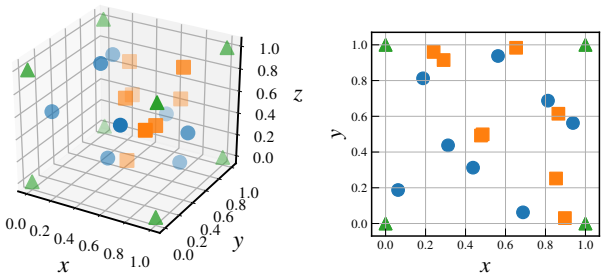

FIG. 2. Example of $\{x, y, z\}$ hyper-parameter locations assigned on an evenly-spaced grid (green triangles), randomly (orange squares), and with Latin hypercube sampling (blue circles), for $M = 8$ training coordinates. A projection of these coordinates into the $\{x, y\}$ plane is shown on the right.

We use the `pyDOE` [48] python module for all simulation designs in this paper. Various sampling options are available, but we choose to maximize the minimum separation between points in hyper-parameter space, while also centering them within the sampling intervals. We compute all simulation coordinates on the unit hypercube, then transform them to the physical hyper-parameter ranges of interest. Figure 2 shows a comparison of how $M = 8$ training coordinates would be assigned in hyper-parameter space according to different simulation design schemes.

B. Data compression

Running population synthesis simulations will provide a catalog of systems, each one with associated measured parameters. In the case of compact binaries, these parameters include component masses, spins, luminosity distance, perhaps eccentricity, etc. A natural way to summarize all this information is to produce histograms of the properties over the entire population; an interpolant could then be used to learn how the input simulation hyper-parameters affect the height of each histogram bin. Although there is nothing formally wrong with this strategy, it misses the opportunity to generate a data-driven basis on which to summarize the parameter distributions, rather than use naive binning. If we simply binned then we would need as many interpolants as bins, which might cause an unnecessary explosion of the computational cost. But if our training distributions lack pathological features, we can form a set of basis distributions that are smaller in number.

To generate a data-driven basis for the simulated distributions of a binary property, we form a data matrix D of shape $N_{\rm bins} \times N_{\rm sims}$. Each column in this matrix corresponds to a single simulation, and contains the normalized bin heights in the histogram for the parameter (flat-

tened over all parameter dimensions, if multi-dimensional histograms are considered), where we a-priori establish a common binning scheme across all simulations. We then use principal component analysis (PCA) [43] on the row-centered matrix to identify a new set of basis distributions:

$$D = U\Sigma V^T, \tag{1}$$

where the magnitude of the singular values along the diagonal of Σ are used to assess the dimensionality of the new basis. We denote $N_{\rm basis}$ as the number of singular values above tolerance that form the restricted $\widetilde{\Sigma}$ diagonal matrix, while the column spaces of U and V are also restricted at $N_{\rm basis}$ to form $\widetilde{U}$ and $\widetilde{V}$. The columns of $\widetilde{U}\widetilde{\Sigma}/\sqrt{N_{\rm basis}}$ are principal components of the parameter distributions that form a natural basis, while columns of $\sqrt{N_{\rm basis}}\widetilde{V}$ correspond to the projection of the original data (bin heights) into the new basis. An interpolant can then be trained on the data in the new compressed basis, such that subsequent predictions are first made in lower dimension before being rotated back into the full-rank binning scheme. Any initial row-centering or scaling is also corrected after a prediction is rotated into full-rank. This data compression scheme identifies characteristic "features" in the parameter distributions.

In the following, the choice of binning scheme (the range and size of bins) is explored case by case. We want to retain the dominant features in our parameter distributions that have sensitivity to hyper-parameters, but also want to avoid an interpolant learning Poisson fluctuations from low occupations bins. Also, for fixed N_{basis}, the compression fidelity may be lower in a finer binning scheme, where the bin heights may fluctuate significantly from Poisson noise.

C. Training an emulator

In regression analysis, or more specifically GW population inference, we need a model to fit to some data. We can assume a parametric form, but we can also be more flexible and let the model be data-driven. In the latter approach, we use the data to train an interpolant which connects the observations by e.g. straight lines (linear interpolation) or low-degree polynomials (spline interpolation). An even more powerful technique than straightforward linear or spline interpolation is GP regression, which treats noisy data as a single random draw from a multivariate Gaussian distribution with a mean vector and covariance function. By optimizing the parameters of a covariance function, and conditioning our predictions of the underlying function on the observations, we let the data tell us the nature of the underlying process rather than enforcing a strict parametric function.

In the rest of this section we define GPs and explain how they can be used as a powerful interpolation tool. There are many excellent treatments of this subject (for general theory see e.g. [49–51]; for ground-based GW applications see [52–54], and for recent applications to Pulsar Timing Arrays see [44, 55]), but here we only summarize the salient points that motivate our work.

1. Gaussian processes

The formal definition of a GP is a (possibly infinite) *"collection of random variables, any finite number of which have a joint Gaussian distribution"* [49]. Instead of parametrizing the underlying function, we are placing a prior (in this case a Gaussian) on the space of possible functions characterized by a mean and covariance. The former is often set to zero and the latter describes how the N points in our sample of the process are correlated [50]. Hence, if we model the underlying process, $f(\mathbf{x})$, as a GP from which our data $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$ are drawn, then formally we can write [49]:

$$\begin{aligned} f(\mathbf{x}) &\sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \\ \mathbf{y} &\sim \mathcal{N}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \end{aligned} \tag{2}$$

where the covariance (or kernel) function is $k(\mathbf{x}, \mathbf{x}') = \langle (f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\rangle$, and as mentioned above we set $m(\mathbf{x}) = \mathbf{0}$.

2. Predictions

We need knowledge of the kernel to constrain the space of possible underlying functions. We train the GP by performing a limited sampling of the underlying process (which in our case are population-synthesis simulations), and condition further predictions on this training data. We account for possible measurement uncertainties on the training data, meaning that we are really measuring noisy values of the underlying process, i.e.

$$\mathbf{y} = f(\mathbf{x}) + \mathbf{n}, \tag{3}$$

where

$$\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \delta(\mathbf{x} - \mathbf{x}'))\,. \tag{4}$$

If we have training data $\mathbf{y}$ measured at $\mathbf{x}$, and we want to predict function values at new points $\mathbf{x}_*$, then we first write the joint distribution of $\mathbf{y}$ and $\mathbf{y}_*$:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K + \sigma_n^2 I & K_*^T \\ K_* & K_{**} \end{bmatrix}\right), \tag{5}$$

where K is the matrix of kernel evaluations over the training data, K_* is the matrix of kernel evaluations between the prediction points and the training data, and K_{**} is the matrix of kernel evaluations over the prediction points.

The conditional distribution of $\mathbf{y}_*$ given $\mathbf{y}$ is [49]

$$\mathbf{y}_* | \mathbf{y} \sim \mathcal{N}(\overline{\mathbf{y}}_*, \mathrm{cov}(\mathbf{y}_*)), \tag{6}$$

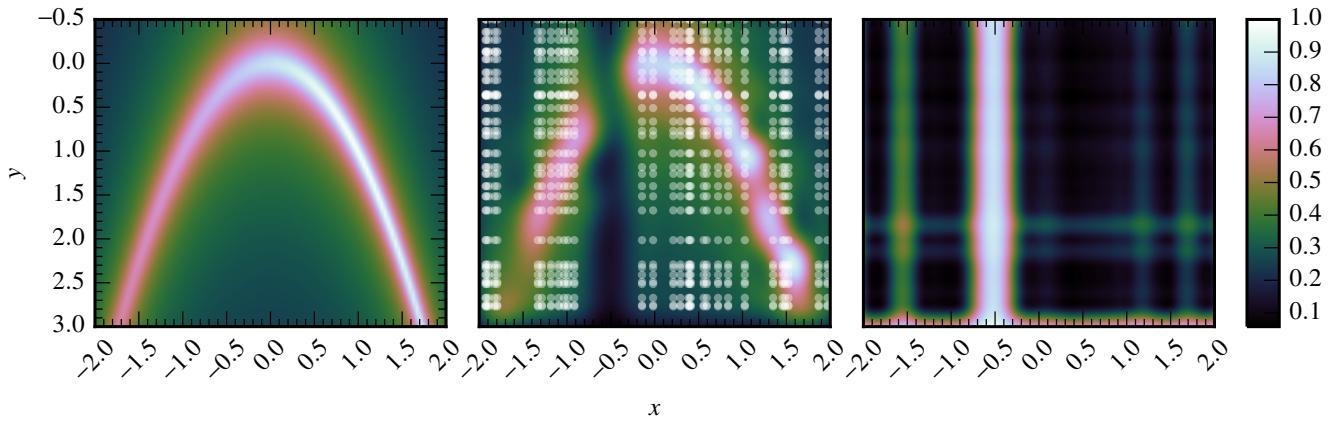

FIG. 3. Training a Gaussian process for prediction. In the left panel we show an inverted offset-Rosenbrock function. In the center panel we show the locations of our training data as white points, along with the GP predicted function values in the background. The right panel shows the uncertainty in the predicted function values of the center panel.

where,

$$\overline{\mathbf{y}}_* = K_*(K + \sigma_n^2 I)^{-1}\mathbf{y}, \tag{7}$$

$$\mathrm{cov}(\mathbf{y}_*) = K_{**} - K_*(K + \sigma_n^2 I)^{-1}K_*^T. \tag{8}$$

Equation ((6)) shows a key result — namely that we have interpolated over our training data by conditioning predictions of new observations on their values. The mean of this conditional distribution $\overline{\mathbf{y}}$ is our prediction, but equally important is the prediction uncertainty $\mathrm{cov}(\mathbf{y}_*)$, which we can propagate through to subsequent inference.

3. *Kernel functions*

The choice of kernel function should be informed by some prior knowledge of the underlying process, but the only formal prerequisite is that it produce a positive-semidefinite covariance matrix. A common choice in the literature is the Squared Exponential (SE) kernel, whose popularity stems from the fact that it is stationary and infinitely differentiable. For training data whose input coordinates are multi-dimensional, this kernel function in a flat metric is:

$$k(x, x') = \sigma_k^2 \exp\left(-\frac{(x_i - x_i')^2}{2\sigma_i^2} - \frac{(x_j - x_j')^2}{2\sigma_j^2} - \dots\right), \tag{9}$$

where each dimension of the input coordinate can have a separate variance $\{\sigma_i^2, \sigma_j^2, \dots\}$, and the kernel has an overall variance scaling σ_k^2. The variance of each dimension acts as a length parameter that dictates the degree with which distant observations can influence each other.

Throughout this paper we use `George` [56], which is a powerful Python library for GP regression. As an example, we sample the following inverted offset-Rosenbrock function at 900 random locations in $[x, y]$ space:

$$g(x, y) = \left[(1 - x)^2 + 100(y - x^2)^2 + 1\right]^{-1/5}. \tag{10}$$

This function is shown in the left panel of Fig. 3, while in the center panel we show the training data locations as white points and the predicted function values in the background. These function values have been predicted by training a GP with an SE kernel. The kernel hyper-parameters were not optimized, but merely set as $\{\sigma_k^2 = 1, \sigma_x^2 = 0.05, \sigma_y^2 = 0.05\}$. The prediction uncertainty is shown in the rightmost panel of Fig. 3, where we see that the predictive model accuracy is worst in the locations where there is a deficit of training data. This feature of GPs is particularly useful since it tells us where in parameter space we must take new samples (i.e. perform new populations synthesis simulations) so that we improve the accuracy of our model. Rather than assume a set of kernel hyper-parameters, we can optimize them; in this case the likelihood (or optimization function) is a Gaussian with an SE kernel, and the training data are treated as a draw from this Gaussian process. We can either map the posterior probability distribution of the kernel hyper-parameters (conditioned on the training data) or simply find the maximum a-posteriori values. In the following, we use MCMC techniques to sample the kernel hyper-parameter posterior distribution, and use the posterior samples to determine the maximum a-posteriori values.

III. INFERENCE TECHNIQUES

In this Section we first outline Bayesian inference as a statistical framework allowing for robust detection and parameter estimation (Sec. III A). We then specify how it is applied to ground-based GW analysis, resulting in

catalogs of measured compact-binary coalescences, each associated with a set of samples drawn from the posterior probability distribution of the event's physical parameters (Sec. III B). Finally, we introduce a hierarchical Bayesian framework for inferring the evolutionary history and progenitor conditions of cataloged GW events, which uses the simulation-trained GP emulator as a parametrized prior (Sec. III C).

A. Bayesian inference

Bayesian inference is a powerful statistical framework allowing models to be robustly tested against data, resulting in probability distributions of the model parameters that are conditioned on both prior expectations and new information [57]. This framework employs Bayes' rule of conditional probabilities, such that the posterior probability of parameters Θ within a model $\mathcal{H}$, implied by data $\mathcal{D}$, is given by:

$$p(\Theta|\mathcal{D},\mathcal{H}) = \frac{p(\mathcal{D}|\Theta,\mathcal{H})p(\Theta|\mathcal{H})}{p(\mathcal{D}|\mathcal{H})}, \tag{11}$$

where $p(\mathcal{D}|\Theta,\mathcal{H}) \equiv \mathcal{L}(\Theta)$ is the likelihood of the model parameters given the data, $p(\Theta|\mathcal{H})$ is the prior probability of the model parameters, and $p(\mathcal{D}|\mathcal{H}) \equiv \mathcal{Z}$ is the fully-marginalized likelihood, or evidence. When inferring credible regions or upper limits for parameters within a single fixed model, the evidence acts as a constant and can be ignored. However it is an important feature for model selection, where the ratio between evidences under different models is known as the Bayes factor. When multiplied by an appropriate prior odds ratio, this becomes the posterior odds ratio, which is essentially the betting odds between the two models.

In parameter estimation we are usually interested in the credible regions for a few parameters. Since Bayesian inference returns probability distributions, we can integrate over over all unwanted nuisance parameters while still incorporating their uncertainty into the measurement spread of parameters that we care about. This technique is known as marginalization. The high-dimensional parameter spaces of models is typically explored using numerical random sampling techniques like Markov Chain Monte Carlo, where the density of the chain samples in parameter space is proportional to the posterior probability density function. As such, all integrations can be trivially tackled through Monte Carlo techniques, e.g.:

$$\int \mathrm{d}x\, f(x)p(x|d,\mathcal{H}) \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i), \tag{12}$$

where $f(x)$ is an arbitrary function, and $p(x|d,\mathcal{H})$ is the posterior probability of x given data d under model $\mathcal{H}$, which we approximate with random samples $i \in [1,\ldots,N]$. We use `emcee` [58] for all sampling in the following.

B. Gravitational-wave parameter estimation

Bayesian inference needs a likelihood function to assess the fitness of the proposed model parameter choices against data, and a measure of the prior probability of these proposed parameters. For ground-based GW analysis, the data is the dimensionless strain computed from the raw interferometric output, which is composed of signal and noise processes. We treat the noise processes as Gaussian and stationary so that we can analytically marginalize over the noise strain, and consider only its power spectral density (PSD), which we assume to be known. For this, we use the Advanced LIGO noise PSD at design sensitivity [59], with a low frequency cutoff at 10 Hz. The strain signal h describing a compact-binary coalescence has 15 parameters: 2 sky-location, 1 polarization angle, 1 initial phase, 3 components of an orbital angular-momentum vector, 2 BH masses, and 2×3 components of the spin vectors. Appropriate sampling of this parameter space will return a set of independent draws from the posterior probability distribution of the signal model. We assume that a catalog of all detected GW events will eventually be issued in the form of sets of these posterior samples (see Refs. [12, 60] for initial steps in this direction).[1]

In the following, we need a simple measure of the detection probability of a compact-binary system. We adopt a frequentist statistic for detection, corresponding to a threshold cut on the expected signal-to-noise ratio (SNR)

$$\rho^2 = 4\int_0^\infty \mathrm{d}f\, \frac{\tilde{h}^*(f)\tilde{h}(f)}{S_n(f)}, \tag{13}$$

where $S_n(f)$ is the one-sided noise PSD, and $\tilde{h}(f)$ is the Fourier-domain waveform. We employ the IMRPhenomD approximant [62] and ignore spins in the SNR calculation, deferring its information content to future work (cf. [63] for possible biases). We access both the Advanced LIGO noise PSD and the waveform approximants through the `pyCBC` python package [64, 65].

A GW signal from a coalescing binary is described by the two polarizations

$$h_+(t) = A(t)\frac{1+\cos^2\iota}{2}\cos\Phi(t), \tag{14}$$

$$h_\times(t) = A(t)\cos\iota\sin\Phi(t), \tag{15}$$

where ι is the binary orbit inclination and all other dependencies are encoded in the signal amplitude $A(t)$ and phase $\Phi(t)$. The response of a (single) detector,

$$h(t) = F_+h_+(t) + F_\times h_\times(t), \tag{16}$$

[1] While this work was being completed, the posterior samples were made available by the LIGO-Virgo Collaboration at dcc.ligo.org/LIGO-T1800235 and Vitale *et al.* [61] for the three events (GW150914, GW151226, LVT151012) in the Advanced LIGO detector's first observing run (O1).

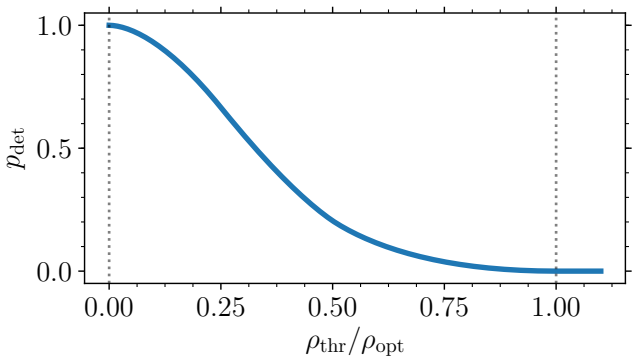

FIG. 4. Detection probability $p_{\rm det}$ as a function of the ratio between a given detection threshold $\rho_{\rm thr}$ and the SNR obtained assuming optimal orientation $\rho_{\rm opt}$. Here we work in the single-detector approximation and assume $\rho_{\rm thr} = 8$.

is modulated by the antenna beam patterns $F_{+,\times}(\theta, \phi, \psi)$, where the three angles describe sky location and polarization content (e.g. Ref. [66]). One can then define the projection parameter [67–69]

$$\omega = \sqrt{\frac{(1+\cos^2\iota)^2}{4} F_+^2(\theta,\phi,\psi) + \cos^2\iota F_\times^2(\theta,\phi,\psi)} \quad (17)$$

and the phase offset

$$\tan\Phi_0 = \frac{2\cos\iota F_\times}{(1+\cos^2\iota)F_+}\,, \quad (18)$$

such that

$$h(t) = A(t)\omega\cos(\Phi(t) - \Phi_0)\,. \quad (19)$$

The parameter ω encapsulate all the angular dependencies of the signal amplitude and satisfies $\max_{\iota,\theta,\phi,\psi}\omega = 1$. From Eq. (13) one thus obtains $\rho = w\rho_{\rm opt}$, where $\rho_{\rm opt}$ is the SNR for an optimally oriented source.

A population synthesis code would typically return a set of binary parameters like masses, spins and distance. The probability that those given binaries exceed a detection threshold $\rho_{\rm thr}$ is computed by averaging over sky location, polarization angle, and inclination. This is equivalent to evaluating the cumulative probability distribution $P(\omega)$ at the ratio between the threshold SNR and the optimal SNR, i.e. $p_{\rm det} = P(\rho_{\rm thr}/\rho_{\rm opt})$. The detectability function is shown in Fig. 4. All of the binary realizations are detectable in the limit $\rho_{\rm opt} \to \infty$, i.e. $p_{\rm det} = 1$. Conversely, none of the realizations are visible below detection threshold, i.e. $p_{\rm det} = 0$ if $\rho_{\rm thr} = \rho_{\rm opt}$. For simplicity we use a single-detector SNR threshold $\rho_{\rm thr} \geq 8$, which has been found to act as a good proxy for more elaborate network analysis [70]. The function $P(\omega)$ is computed with a Monte Carlo as implemented in the python package `gwdet` [71].

C. Hierarchical population inference

1. *Priors and hyper-parameters*

Choices of parameter priors may be motivated by underlying physical intuition (e.g. neutron star masses can not be greater than $\sim 4M_\odot$) or fundamental constraints (e.g. masses should be positive, speeds can not exceed the speed of light, etc.). However, sometimes intuition or fundamental constraints do not lead us to a definitive prior, as in the case of the astrophysical distribution of compact object masses and spins. In some cases one might be able to make a reasonable guess at the form of the distribution (e.g. Gaussian), but the mean and width may be unknown. Or perhaps even the form itself is completely unknown, and only dictated by unknown properties of the progenitor system. In this case, we can extend our model to an additional level (hence *hierarchical* inference) by using a parametrized prior. The parameters of these priors are the hyper-parameters, and they themselves will have hyper-priors.

2. *Likelihoods and posteriors*

Hierarchical inference is discussed in detail elsewhere (e.g. Refs. [27, 72–77]), but we summarize the salient points here. We make specific use of the formalism in Mandel *et al.* [78] and Farr *et al.* [79]. The goal is to simultaneously infer the joint posterior probability distribution of the measured physical parameters of each event, as well as the hyper-parameters describing the statistical properties of the entire population.

The joint probability of strain data from all GW signals $\{h_k\}$ (where $k \in [1, \ldots, N]$ indexes each event), and associated physical parameters describing each signal θ_k is

$$p(\{h_k\},\{\theta_k\}|\beta) = p(\{h_k\}|\{\theta_k\})\,p(\{\theta_k\}|\beta), \quad (20)$$

where β are the population hyper-parameters. The GW signals will be produced at a certain rate in parameter space as a function of the hyper-parameters. We first consider a discrete representation of the physical parameter space (e.g. masses, redshits, etc) divided into bins, $l \in [1, \ldots, N_l]$. The data are then the number of events detected in a given bin in this parameter space n_l. Assuming non-overlapping statistically-independent signals (and thus bins)[2], the likelihood is the product of a Poisson process in each bin:

$$p(\{n_l\}|\beta) = \prod_{l=1}^{N_l} \frac{(r_l(\beta))^{n_l} e^{-r_l(\beta)}}{n_l!}, \quad (21)$$

[2] This assumption is expected to fail for future 3rd-generation ground-based detectors, as well as the LISA space mission.

where $r_l(\beta)$ is the expected rate of events in bin l as function of hyper-parameters β. If we make the bins infinitesimally small, then each bin will either have 1 or 0 events. This gives the continuum limit

$$p(\{\theta_k\}|\beta) \propto e^{-N_\beta} \prod_{k=1}^{N} r(\theta_k|\beta), \tag{22}$$

where $N_\beta = \iint dh\, d\theta\, p(h|\theta)\, r(\theta|\beta)$ is the expected total number of events for a population with hyper-parameters β, and $r(\theta|\beta) = N_\beta\, p(\theta|\beta)$ such that $\int d\theta\, p(\theta|\beta) = 1$. The likelihood $p(h|\theta)$ is normalized over the data, so while the data integral is trivial here we will see soon why its explicit marginalization is useful. Plugging Eq. (22) into Eq. (20), and again using the statistical independence of signals, gives

$$p(\{h_k\},\{\theta_k\}|\beta) \propto e^{-N_\beta} \prod_{k=1}^{N} p(h_k|\theta_k)\, r(\theta_k|\beta). \tag{23}$$

The measured data are usually thresholded using a detection statistic to decide which signals are robust events, and which are spurious or untrustworthy. Upon examining the data, we partition N into "observable" ($N_{\rm obs}$) and "non-observable" ($N_{\rm nobs}$), so that Eq. (23) becomes

$$p(\{h_i\},\{\theta_i\},\{h_j\},\{\theta_j\}|\beta) \propto e^{-N_\beta} \left[\prod_{i=1}^{N_{\rm obs}} p(h_i|\theta_i)\, r(\theta_i|\beta)\right] \left[\prod_{j=1}^{N_{\rm nobs}} p(h_j|\theta_j)\, r(\theta_j|\beta)\right]. \tag{24}$$

We now marginalize over the data and parameters of the non-observable events, and divide the probability by $N_{\rm nobs}!$ to mitigate over-counting through marginalization. We also marginalize over the number of non-observable events, $N_{\rm nobs}$, from 0 to ∞:

$$\begin{aligned} p(\{h_i\},\{\theta_i\}|\beta) &\propto e^{-N_\beta} \left[\prod_{i=1}^{N_{\rm obs}} p(h_i|\theta_i)\, r(\theta_i|\beta)\right] \sum_{N_{\rm nobs}=0}^{\infty} \frac{(N_\beta^{\rm ndet})^{N_{\rm nobs}}}{N_{\rm nobs}!} \\ &\propto e^{(N_\beta^{\rm ndet} - N_\beta)} \prod_{i=1}^{N_{\rm obs}} p(h_i|\theta_i)\, r(\theta_i|\beta) \\ &\propto N_\beta^{N_{\rm obs}}\, e^{-N_\beta^{\rm det}} \prod_{i=1}^{N_{\rm obs}} p(h_i|\theta_i)\, p(\theta_i|\beta), \end{aligned} \tag{25}$$

where,

$$\begin{aligned} N_\beta^{\rm det} &= \int\int_{\{h\in[\rm detection]\}} dh d\theta\, p(h|\theta)\, r(\theta|\beta) \\ &= \int d\theta\, p_{\rm det}(\theta)\, r(\theta|\beta) \\ &= N_\beta \times \int d\theta\, p_{\rm det}(\theta)\, p(\theta|\beta) \\ &= N_\beta \times \epsilon_\beta \end{aligned} \tag{26}$$

is the expected number of detected events in a population model with hyper-parameters β, such that $N_\beta = N_\beta^{\rm det} + N_\beta^{\rm ndet}$. The probability of detection as a function of binary parameters is given by $p_{\rm det}(\theta)$ from Sec. III B. The efficiency $\epsilon_\beta = \int d\theta\, p_{\rm det}(\theta)\, p(\theta|\beta)$ denotes the fraction of merging systems that are detectable for a given hyper-parameter coordinate.

Equation (25) is appropriate if we fully model all factors influencing the number and distribution of detectable GW events, such as the local merger-rate density, the duty cycle of the detectors, etc. In our analysis we construct $r(\theta|\beta)$ from population synthesis simulations, from which we record the fraction of initialized stars that were evolved to become merging BH-BH systems. We do not want to make our analysis sensitive to duty-cycle choices or poorly-constrained scaling parameters that could affect rates, so we marginalize over such factors [27, 38, 76, 80]. This is done by marginalizing over N_β with the prior $p(N_\beta) \propto 1/N_\beta$, such that [81]

$$p(\{h_i\},\{\theta_i\}|\beta) \propto (N_{\rm obs}-1)! \prod_{i=1}^{N_{\rm obs}} \frac{p(h_i|\theta_i)\, p(\theta_i|\beta)}{\epsilon_\beta}. \tag{27}$$

The first term in the numerator is the single-event likelihood used for GW parameter-estimation. We do not want to repeat all of the effort that went into reducing the raw detector output to a set of likelihood evaluations. Rather, we assume a GW catalog will eventually be provided in the form of a set of posterior samples for each event:

$$p(\theta_i|h_i,\overline{\beta}) = \frac{p(h_i|\theta_i)p(\theta_i|\overline{\beta})}{p(h_i|\overline{\beta})}, \tag{28}$$

where $\overline{\beta}$ denotes the prior for the BH/NS parameters chosen by the issuers of the catalog (e.g. uniform in component masses, comoving volume, etc.). Plugging Eq. (28) into Eq. (25), and Monte Carlo integrating over the posterior distribution of event parameters with Eq. (12) gives

$$p(\{h_i\}|\beta) \propto Z_{\overline{\beta}} \times N_\beta^{N_{\rm obs}}\, e^{-N_\beta \epsilon_\beta} \prod_{i=1}^{N_{\rm obs}} \left\langle \frac{p(\theta_i|\beta)}{p(\theta_i|\overline{\beta})} \right\rangle_{{\rm post},i}, \tag{29}$$

where $Z_{\overline{\beta}}$ is the evidence for the interim prior model using the data from all observed events. This is a constant and can thus be ignored. The expectation value in Eq. (29) is taken over samples drawn from the joint posterior distribution of each event in the GW catalog, while the argument is the ratio of the rate of detected-event parameters under our new parametrized model (constructed from simulations) versus the interim prior (used in the catalog construction). Dividing out the influence of the interim prior used in the original event analysis is crucial (e.g. Ref. [61]), since our goal is to re-analyze the entire catalog under the new parametrized prior that has been constructed from simulations. For the examples reported in this paper, we approximate the interim prior as being

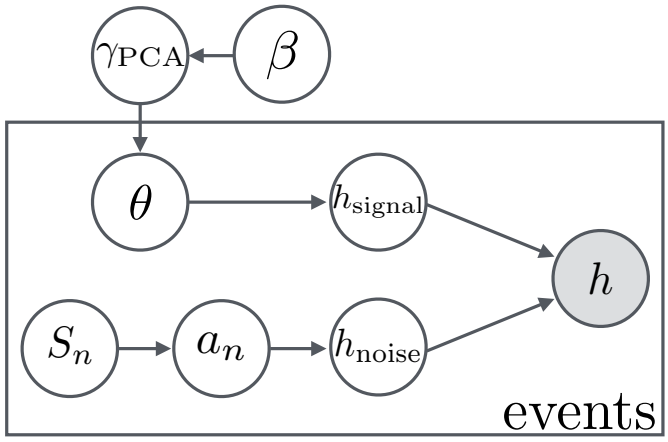

FIG. 5. A probabilistic graphical model illustrating Eq. (20). The detector output, h, depends on noise and signal processes. The noise may be decomposed onto a Fourier basis with coefficients, a_n, whose variance in turn may be constrained by a model for the power-spectral density, S_n. The strain induced by each signal depends on the intrinsic and extrinsic parameters of each binary θ. We place a parametrized prior on a subset of these parameters, given by orthogonal basis distributions determined from PCA of population synthesis simulations, $\gamma_{\rm PCA}$. The amplitude of each basis distribution has a Gaussian prior from GP training on these simulations, informed by some hyper-parameters, β.

uniform over the region of parameter space with likelihood support, so that we can safely ignore this subtlety.

Monte Carlo integrating over the posterior distribution of event parameters in Eq. (27) gives

$$p(\{h_i\}|\beta) \propto Z_{\overline{\beta}} \times (N_{\rm obs}-1)! \prod_{i=1}^{N_{\rm obs}} \frac{1}{\epsilon_\beta} \left\langle \frac{p(\theta_i|\beta)}{p(\theta_i|\overline{\beta})} \right\rangle_{{\rm post},i}. \tag{30}$$

The rate, N_β, and distribution, $p(\theta|\beta)$, are constructed using the simulation and emulation scheme described in Sec. II, where the former is found by training on the fraction of ZAMS stars that form merging BH-BH systems. Fig. 5 shows the probabilistic graphical model for our inference framework, and illustrates the chain of conditional dependencies for constraining the parameters of each event with a prior that is a function of progenitor and evolutionary properties. We use both Eq. (29) and Eq. (30) in the following test cases.

IV. RESULTS

We now implement our new framework on three case studies. These case studies begin with a toy model (Sec. IV A), then increase in complexity and astrophysical realism using both public data (Sec. IV B) and tailored simulations (Sec. IV C) to showcase how one might use our findings in practice.

A. Toy Model

Our first demonstration corresponds to the inference of binary spin-alignment distributions. Spin alignments are indeed recognized as one of the cleanest indicators for constraining BH formation and evolutionary processes [30, 39, 82–88]. Here we implement the approach developed by Talbot and Thrane [82]. The observed quantities in this model are the projection of each binary component's spin onto the orbital angular momentum vector:

$$z_1 = \hat{L}\cdot\hat{S}_1, \quad z_2 = \hat{L}\cdot\hat{S}_2, \tag{31}$$

where $z_{\{1,2\}} \in [-1,1]$. Dynamical capture mechanisms in, e.g., a globular cluster are expected to produce an isotropic distribution of spin alignments

$$p_0(z_1,z_2) = \frac{1}{4}. \tag{32}$$

For field binaries, the evolutionary path of each progenitor star (in particular natal kicks during supernova) is assumed to produce a truncated Gaussian distribution of alignments. Two hyper-parameters σ_1 and σ_2 control the degree with which (anti-)alignment is favored:

$$p_1(z_1,z_2) = \frac{2}{\pi}\frac{1}{\sigma_1}\frac{e^{-(z_1-1)^2/2\sigma_1^2}}{{\rm erf}(\sqrt{2}\sigma_1)}\frac{1}{\sigma_2}\frac{e^{-(z_2-1)^2/2\sigma_2^2}}{{\rm erf}(\sqrt{2}\sigma_2)}. \tag{33}$$

In this model, $\sigma = 0$ produces perfect alignment, while $\sigma = \infty$ tends to the dynamic-capture distribution.

We use $p_1(z_1,z_2)$ as the test destribution to be inferred. This probability function has hard-edges at $[z_1 = \pm 1, z_2 = \pm 1]$, making it challenging to learn and thus an excellent testbed to test our framework. The observed parameters from each GW binary event are $\theta \in \{z_1, z_2\}$, and the hyper-parameters of the population are $\beta \in \{\sigma_1, \sigma_2\}$. The parameter probabilities are represented on a 40×40 binning in $\{z_1, z_2\}$ space.

We generate training data using Eq. (33) for a range of $\sigma_{1,2} \in [0.1, 10]$ values, sampled uniformly in log-space. To examine how many training datasets are needed, we create grids of training data with different densities in hyper-parameter (i.e. β) space. We find a compressed basis representation of the training-data distributions, then train a GP at each bin in the compressed parameter space. In all cases we find that the initial parameter binning can be compressed by a factor of ~ 500 with high-fidelity[3]. In this case the compression and training is performed on the logarithm of the training data, since this reduces the dynamic range of values across parameter space and ensures that the predicted proability values

[3] We compute the normalized inner product of the training data (flattened to be the vector of all samples in the dataset) with the compressed data (which has been rotated back into the full parameter basis). With only 3 reduced basis distributions, corresponding to a compression of $(40\times 40)/3 \approx 533$, we achieve discrepancies from true that are of $\mathcal{O}(10^{-16})$.

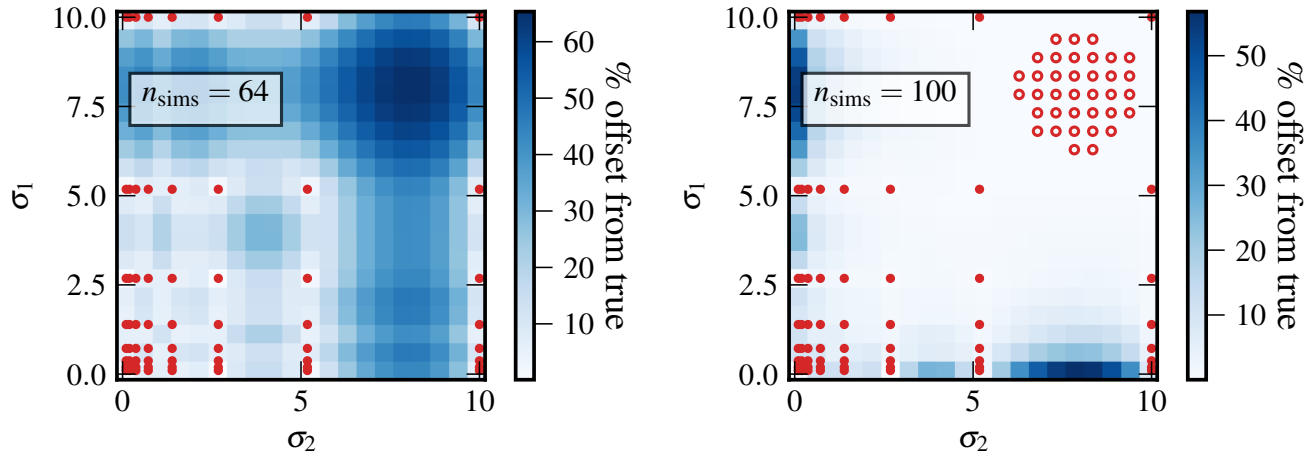

FIG. 6. Testing the accuracy of our GP emulator for the model of Eq. (33). In the left panel we create training data on an evenly-spaced 8×8 grid in $\log_{10} \sigma_{1,2}$ space (red points). We achieve a data compression factor of ~ 500, then train a GP in each of the reduced basis features. The GP prediction is compared to the analytic result across $\sigma_{1,2}$ space by taking the GP-mean (offset by 1 σ), rotating back to the full $z_{1,2}$ basis, then finding the maximum difference from the analytic value in any $z_{1,2}$ bin. Low accuracy locations are used to inform the positions at which new simulations are performed. These additional points are shown in the right panel as empty circles, where we see that their addition improves accuracy across the entire hyper-parameter space.

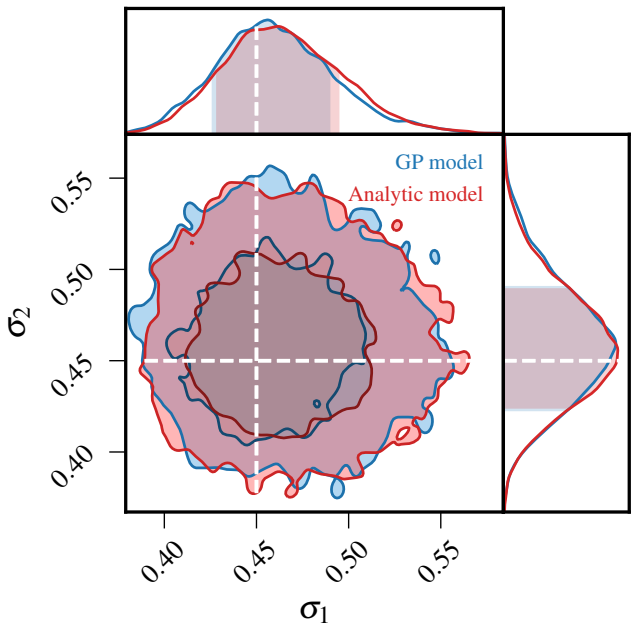

FIG. 7. Comparison of posterior recoveries of population hyper-parameters from a catalog of 100 sources with spin-alignment distribution given by Eq. (33) [82]. The true hyper-parameter coordinate, $\{\sigma_1 = 0.45, \sigma_2 = 0.45\}$ is indicated via intersecting white dashed lines.

will always be positive. We can now predict the distribution values in compressed parameter space, and rotate this back into the full parameter space to construct the final predictions.

Figure 6 shows validation studies for different numbers of initial training data. For an evenly-spaced grid of $8 \times 8 = 64$ training datasets in hyper-parameter space, we achieve an accuracy of better than $\sim 50\%$ across the majority of the space. The worst performance occurs in parts of hyper-parameter space that are voids of simulations. We find the 36 worst accuracy locations, and add these as additional simulations to improve accuracy to better than 10%. Similar accuracy is given by an Latin-hypercube design of 100 training datasets.

We now test our framework on a simulated population, consisting of 100 sources drawn from $p(z_1, z_2)$ with $\beta = \{\sigma_1 = 0.45, \sigma_2 = 0.45\}$. A comparison of the joint posterior probability distribution of $\{\sigma_1, \sigma_2\}$ as recovered by the analytic model [Eq. (33)] and the GP framework is shown in Fig. 7. The GP framework is trained on 100 simulations from a Latin-hypercube design; we use this design because it is our standard approach for efficiently sampling the high-dimensional hyper-parameter space of binary stellar evolution, and it gives similar emulation accuracy to the adaptive design in the right panel of Fig. 6. In this analysis, we have propagated all uncertainties from the GP prediction and the hyper-parameters of the trained GP covariance function into the final model. The agreement is excellent, with the true hyper-parameter coordinate lying well within the 68% credible region of both techniques. We have not incorporated the effect of indi-

vidual event measurement uncertainties, which will be explored in the next examples.

B. COMPAS Populations

We now test our framework on an example with greater astrophysical realism. We take publicly-available populations[4] of synthesized binary BHs from Stevenson *et al.* [25] as training data. In the aforementioned paper, the authors introduce `COMPAS`: a code (broadly similar to `BSE`) for evolving zero-age-main-sequence (ZAMS) binary star systems through classical isolated evolution (i.e. including common-envelope stages). By simulating low-metallicity populations and following the binary evolution, the authors find that all three initial Advanced LIGO events (GW150914, GW151226, and LVT151012) could have been formed with a single model in an environment with $Z \sim 0.05 Z_\odot$. In Ref. [25], the statistic for checking whether a given simulated binary was consistent with forming each individual detected GW event was whether the simulated binary's total mass (chirp mass) fell within the quoted 90% credible region for GW150914 (GW151226, LVT151012), and whether the mass-ratio exceeded the quoted 90% credible lower bound. While this is a reasonable measure of consistency, it does not provide a corresponding measure of statistical credibility for the inferred progenitor metallicities. By contrast, our framework allows the posterior probability distribution of progenitor metallicities to be recovered.

We use populations produced with fiducial assumptions under different metallicities, corresponding to $Z = \{0.05, 0.1, 0.25\} Z_\odot$. In this example, Z is the only hyper-parameter that we aim to infer. All binaries reported merge within a Hubble time, and we incorporate detector selection effects using the detection probability mentioned in Sec. III B. In principle we would use the binary component masses, spin information, and redshift to discriminate progenitor properties and evolutionary paths. But since there is only a limited amount of information that can be inferred based on these three training populations, we opt for simplicity and only use the chirp mass information from each binary. We do not consider rate information either, such that our likelihood is given by Eq. (30). By using these publicly-available populations as training data, we implicitly approximate all BH systems as forming from progenitors with a common metallicity.

We compress histograms of each population's chirp masses from 80 initial bins down to a PCA basis of size 2 (which is set by the small number of training populations). The compressed training data is then interpolated over metallicity using a GP with a squared-exponential kernel. This procedure gives a model for the distribution of detectable chirp masses as a function of metallicity.

[4] Populations available at http://www.sr.bham.ac.uk/compas/data.

TABLE I. The existing catalog of binary BH detections from Advanced-LIGO–Advanced-Virgo, with measured source-frame chirp masses and merger redshifts reported as median values and associated 90% credible bounds.

Event	Chirp mass $\mathcal{M}$	Merger redshift z	Refs.
GW150914	$28.1^{+1.8}_{-1.5}\,M_\odot$	$0.09^{+0.029}_{-0.036}$	[2, 89]
LVT151012	$15.1^{+1.4}_{-1.1}\,M_\odot$	$0.201^{+0.086}_{-0.091}$	[2]
GW151226	$8.88^{+0.33}_{-0.28}\,M_\odot$	$0.094^{+0.035}_{-0.039}$	[2, 3]
GW170104	$21.1^{+2.4}_{-2.7}\,M_\odot$	$0.18^{+0.08}_{-0.07}$	[4]
GW170608	$7.9^{+0.2}_{-0.2}\,M_\odot$	$0.07^{+0.03}_{-0.03}$	[5]
GW170814	$24.1^{+1.4}_{-1.1}\,M_\odot$	$0.11^{+0.03}_{-0.04}$	[6]

We perform a simple test using chirp-mass and redshift information from the catalog of existing BH detections, see Table I. We make the very simple approximation that the source-frame chirp mass and merger redshift posterior distributions are Gaussian and uncorrelated, from which we can easily draw posterior samples. We draw 100 independent posterior samples for each event and use these samples to propagate parameter-estimation uncertainty into our population hyper-parameter inference. This is obviously a highly simplified representation of the real event posteriors, but it outlines the scheme one would use when provided with the samples from the true GW catalog.

Another subtlety that we do not consider here (but that must be accounted for in a real analysis) is the influence of the original priors from the parameter-estimation analysis of each individual event (c.f. Sec. III C). In the current Advanced-LIGO–Advanced-Virgo searches, the component mass priors are uniform, while the luminosity distance prior assumes the mergers occur uniformly in comoving volume. These choices do not translate to uniform priors in chirp mass or redshift, so that we should re-weight the posterior samples from each event to reflect the likelihood, then apply our newly-formulated parameter priors (as a function of population hyper-parameters) to the entire detected event catalog. In this analysis, we simply assume that the chirp mass and redshift priors were uniform in the analysis of each GW event.

The resulting posterior distribution for progenitor metallicity is shown Fig. 8, where the 68% and 90% upper limits are found to be $Z < 0.12\,Z_\odot$ and $Z < 0.16\,Z_\odot$, respectively. This is in broad agreement with Stevenson *et al.* [25], who found that the three events required $Z \simeq 0.05\,Z_\odot$. Our constraints reflect uncertainties in the GP model prediction and the parameter estimation of each event. In Fig. 9 we also show the reconstructed intrinsic chirp-mass distribution of binary BHs at metallicities corresponding to our credible limits, as well as the original training distributions. We see that our model correctly interpolates the physical behavior on which it

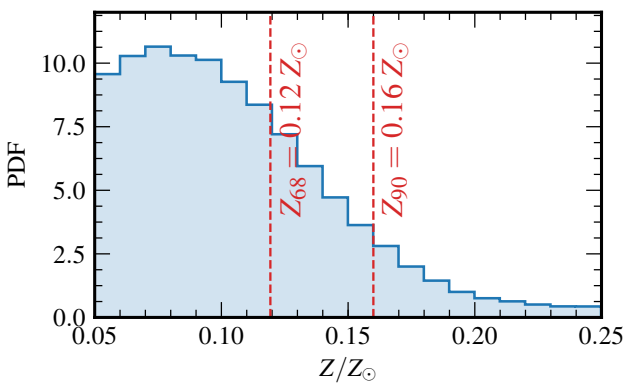

FIG. 8. Posterior probability distribution of progenitor metallicity Z, as inferred by an analysis of the current BH catalog in Table I using a model for the chirp mass distribution that is conditioned on simulations from [25]. Dashed vertical lines marks the 68% and 90% confidence intervals.

was trained (including some sharp features), namely that the distribution of chirp masses shifts to smaller values as the progenitor metallicity is increased. Physically, this is because stellar winds are weaker in stars with lower metallicity, that thus tend to form heavier BHs like the ones detected by Advanced LIGO [25, 34–36]. The events of the current binary BH catalog are shown as vertical bands corresponding to the 90% credible region of chirp mass.

C. BSE Population Synthesis

To further showcase the effectiveness of our statistical framework, we now consider a more elaborate set of input data. We perform a dedicated program of population-synthesis simulations to predict properties of BH binaries from isolated binary stars.

We use a modified version of the public population synthesis code `BSE` [18, 90]. The modifications implemented here are the same described in Refs. [36, 91]: wind mass loss prescriptions according to Ref. [92] and core-collapse remnant mass relationship following Ref. [20]. These minimal updates are necessary to generate any BHs of masses $\gtrsim 10M_\odot$ like the ones that are now detected, and thus to attempt a comparison with the Advanced-LIGO–Advanced-Virgo data. We stress, however, that this study is not meant to rival with the full complexity of state-of-the-art binary evolution codes, but rather highlight the potential of our inference pipeline.

`BSE` requires us to specify distributions of binary stars on their zero-age main sequence (ZAMS), and a variety of flags encoding assumptions of the underlying stellar physics. We distribute primary masses m_1 from an initial mass function $p(m_1) \propto m_1^{-2.3}$ in $[5, 100]M_\odot$; mass ratios $q = m_2/m_1$ uniformly in $[0, 1]$; initial separations R uniformly in $\log_{10}$ in $[10, 10^5]R_\odot$; eccentricities e from a thermal distribution $p(e) \propto e$; and redshifts z uniformly in comoving volume using the Plank cosmology [93] (c.f. Ref. [29] for similar choices).

The evolutionary flags are the quantities that should be treated as hyper-parameters, and that could potentially be constrained with current and future catalogs of GW events. For simplicity, we present results considering a 3-dimensional hyper-parameter space, but our method is fully generalizable and scalable to higher dimensions. We fix all flags to their default value in BSE, except for the following three:

1. *Metallicity of the ZAMS star: Z.* As already highlighted above, the progenitor metallicity has a large impact on the properties of the resulting BHs. Metallicity strongly affects massive star winds and thus the mass that remains available to form the final compact object [22, 24, 92, 94–97]. Here we consider a metallicity range $0.0001 \leq Z \leq 0.03$ where $Z_\odot = 0.02$ [18].

2. *Kicks imparted to BHs at formation: $\sigma_{\rm k}$.* As stars collapse (perhaps exploding into supernovae), asymmetries in the emitted material and neutrinos may impart a recoil to the newly formed compact object (e.g. Ref. [98]). Observations of galactic pulsar proper motions suggest that NS recoils are well modeled by a single Maxwellian distribution with 1D root-mean-square $\sigma_{\rm k} \sim 265$ km/s [99, 100]. Whether BHs receive any kick at formation is still a matter of debate. On the one hand, X-ray binary measurements hint at large kick velocities [101] (c.f. also Ref. [102] for a GW constraint). Conversely, theoretical arguments and simulations suggest that kicks for BHs might be suppressed because of material falling back after the explosion [98, 103, 104]. This is a clear case where a method like ours, allowing for a direct estimate of $\sigma_{\rm k}$, might show its potential. We consider BH recoils in the range $0\,{\rm km/s} \leq \sigma_{\rm k} \leq 265\,{\rm km/s}$ independently of BH mass or other parameters (see Ref. [40] for a discussion of this point).

3. *Efficiency of the common envelope: $\alpha_{\rm ce}$.* After the first star collapses, the binary system consists of a BH and an extended star. As this second star expands into a supergiant, it may overflow its Roche Lobe and undergo unstable mass transfer to the BH [105–108]. The envelope of the giant engulfs the companion BH. In this process, known as the common-envelope stage, a fraction $\alpha_{\rm ce}$ of the binary's orbital energy is transferred to the envelope, thus hardening the binary. In the standard evolutionary channel considered here, common envelope evolution is the key stage to produce BHs able to merge within a Hubble time. The details of the common envelope phase are still very uncertain [109–112], and are arguably one of the most

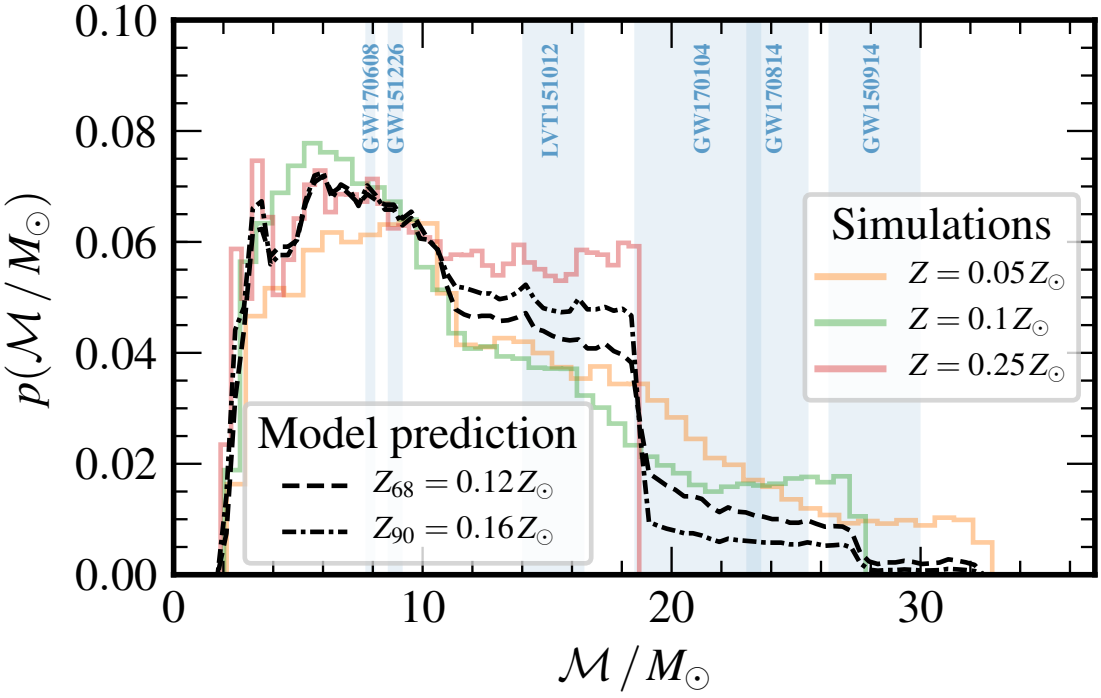

FIG. 9. Intrinsic distribution of BH binary chirp masses for progenitor metallicity values corresponding the simulations by [25] (colored lines) and the 68% and 90% upper limits from an analysis of the current GW catalog (black dashed lines). The chirp masses of the GW events in the catalog are shown with vertical blue bands.

important stellar (hyper-)parameters that can potentially be measured with GW data. Here we vary $\alpha_{\rm ce}$ in $[0.001, 10.0]$.

We use $\{Z, \sigma_{\rm k}, \alpha_{\rm ce}\}$ as hyper-parameters, thus implicitly assuming that all stars in the same simulated universe share common values of those quantities. While this might be a good working assumption for, e.g., $\alpha_{\rm ce}$, it is surely not true for other parameters like the metallicity. That said, our methods can be straightforwardly generalized to a distribution of metallicities with parameters that can be treated as hyper-parameters in our inference instead of Z itself (much like $\sigma_{\rm k}$, which is a parameter in the Maxwellian kick distribution, not the kick velocity itself).

We perform 125 BSE simulations distributing $\log_{10} Z$, $\sigma_{\rm k}$, and $\log_{10} \alpha_{\rm ce}$ on a Latin hyper-cube as described in Sec. II A and drawing $N = 10^7$ ZAMS binaries at each point in hyper-parameter space. Each of these $125 \times N$ simulated stars is filtered according to two criteria: (i) a BH binary is formed, and (ii) it merges before $z = 0$. Binaries passing these cuts are assigned an Advanced LIGO detection probability, $p_{\rm det}$ (c.f. Sec. III B).

Each BSE simulation returns a population of BH binaries characterized by their masses and merger redshifts, which we use as measured event parameters in our statistical inference. Examples of the intrinsic $\{\mathcal{M}, z\}$ distribution for two of these simulations are shown in Fig. 10, where low Z values ensure stars are able to form massive BHs. The relative merger rate (i.e. the fraction of ZAMS stars that form merging BH binaries) is shown

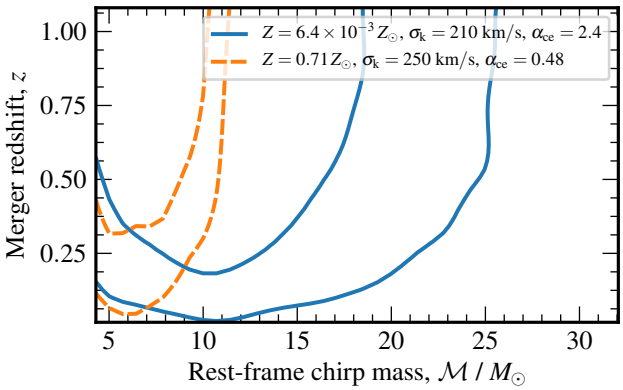

FIG. 10. An example of two BSE training simulations, showing the intrinsic $\{\mathcal{M}, z\}$ distribution of merging BH binaries. Contours enclose 68% and 90% of simulated binaries, where the blue solid lines are for a very low metallicity progenitor scenario, while the orange dashed lines are for a simulation close to solar metallicity.

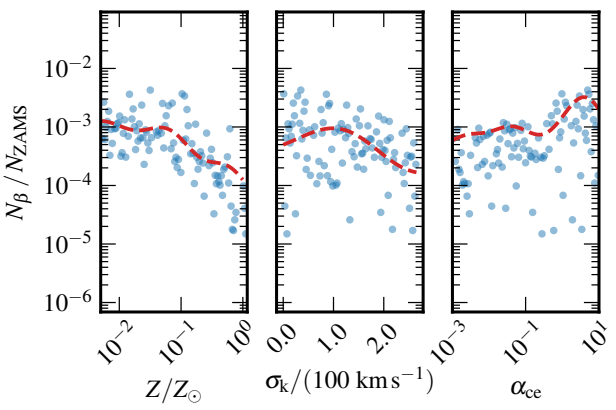

FIG. 11. Fraction of ZAMS stars that form merging binary BH systems. The three panels show fractions in each of our three hyper-parameters: metallicity Z, natal kicks $\sigma_{\rm k}$ and common-envelope efficiency $\alpha_{\rm ce}$. The dashed lines in each panel show predictions from a GP that has been trained on these rates, with only the hyper-parameter relevant to the panel varied in the prediction.

in Fig. 11. The rate decreases with Z because (i) fewer BHs are formed in favor of NSs (which are not considered here for simplicity) and (ii) stars become puffier at large Z and are more likely to merge earlier in the evolution (e.g. Ref. [36]). The rate also decreases with $\sigma_{\rm k}$ because strong kicks more easily unbind binaries (e.g. Ref. [40, 113]).

We do not know a-priori how many ZAMS stars survive as merging BH binaries. Some points in hyper-parameter space lead to only a handful of events, giving a jagged distribution in parameter space that suffered from finiteness. To counter this, we require a simulation to provide at least 500 systems in order to be included in our training and validation procedures. This leaves 115 out of the original 125 hyper-parameter coordinates. Even though this renders our simulation coordinates no longer a perfect LH design, it creates a training set with smoother and more robust parameter distributions. Out of the surviving 115 simulations, we train our GP emulator on a randomly chosen 100, with another 14 selected for independent validation of the GP, and the final simulation left as a test population for the full hierarchical Bayesian pipeline.

For each training simulation, we create a normalized KDE-smoothed[5] 2D distribution in intrinsic chirp mass, $\mathcal{M}$, and merger redshift, z, with a common 20×20 binning scheme. The distributions are PCA-compressed by a factor of 8, with a compression fidelity of better than 0.01%. The remaining 50 features (or "bins") in the compressed distributions are each interpolated over the three-dimensional hyper-parameter space of $\beta = \{\log_{10} Z, \sigma_{\rm k}, \log_{10} \alpha_{\rm ce}\}$ using GPs with squared-exponential kernels. We denote a match statistic that is the normalized inner product of the bin heights in the validation distribution with the GP-predicted distribution. With maximum a-posteriori GP kernel parameters, the 14 validation distributions (KDE-smoothed and normalized) all match their GP-predicted distributions to better than 7%. We also train a separate GP on the fraction of ZAMS stars that survive as merging binary BH systems, which was used to make the smooth rate curves in Fig. 11. This is convolved with detector selection effects to compute the fraction of merging systems that are detectable in Advanced LIGO.

We still have one population that was held out of the GP emulator training and validation, which we now use as data for a test of the entire hierarchical Bayesian pipeline. The hyper-parameters of this population are $\beta = \{Z = 7.3 \times 10^{-4}, \sigma_{\rm k} = 100\,{\rm km/s}, \alpha_{\rm ce} = 0.021\}$. We weight each system in the population by its detection probably, then randomly select 100 to be our catalog, corresponding to (depending on duty-cycle and sensitivity assumptions) a few years of Advanced Advanced-LIGO–Advanced-Virgo observations. The evaluated match statistic between this distribution and our GP prediction is $\sim 0.5\%$. We take two approaches to analyze this catalog:

(i) using only the information given by the $\{\mathcal{M}, z\}$ distribution of sources, see Eq. (30);

(ii) artificially scaling the rate-GP to predict 100 detected events for the test hyper-parameters so that we can use a Poisson likelihood, see Eq. (29).

The recovered posterior probability distributions of population hyper-parameters are shown in Fig. 12, where all are consistent with the true values. We have not marginalized over GP kernel posteriors or the GP prediction uncertainties so that we may see the effect (or in this case lack thereof) of systematic offsets from interpolation errors. We have also not modeled parameter uncertainties in the cataloged events, but these can be straightforwardly incorporated.

As a final test, we analyze the current Advanced-LIGO–Advanced-Virgo catalog from Table I, following the same assumptions as in Sec. IV B. We use Eq. (30), and marginalize over all cataloged event parameter uncertainties. As expected, with only 6 events the posterior distributions for $\sigma_{\rm k}$ and $\alpha_{\rm ce}$ are broad and do not significantly update their priors. However, we place a constraint on progenitor metallicity corresponding to $Z < 0.09\,Z_\odot$ at 90% credibility. The marginalized mass and redshift distribution of binary BH mergers for the maximum a-posterioi hyper-parameters from this analysis are shown in Fig. 13. We stress again that all constraints are subject to our assumptions and minimal updates of `BSE`, which are only intended to show the capabilities of our approach.

[5] We use the `scipy.stats` implementation of Gaussian kernel density estimation with a bandwidth selected by Scott's Rule [114].

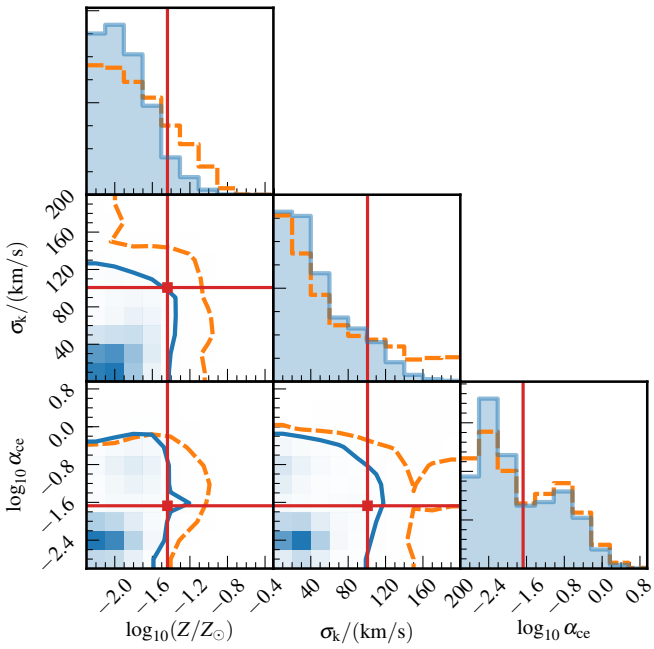

FIG. 12. Corner plot showing 1D-marginalized posterior distributions of binary BH population hyper-parameters along the diagonal, and pairwise 2D-marginalized posterior distributions in the lower axes (lines denote 90% credible regions). The true hyper-parameters are indicated with red lines. The data were 100 binary BHs from a population simulated with `BSE` that was held out of our GP emulator training. Results are for a distribution-only likelihood [*orange dashed*, Eq. (30)], and a re-scaled Poisson-rate likelihood [*blue solid*, Eq. (29)].

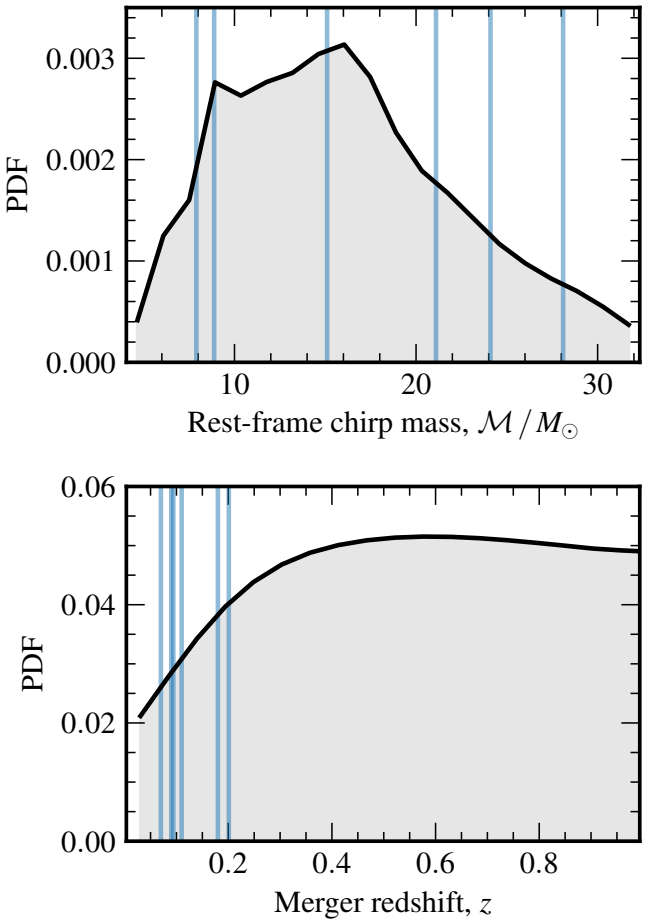

FIG. 13. Marginalized binary BH population distributions of rest-frame chirp mass and redshift for the maximum a-posteriori hyper-parameters from an analysis of the current Advanced-LIGO–Advanced-Virgo catalog. These are the intrinsic merger distributions, rather than convolved with detector selection effects. The blue vertical lines indicate the parameters of cataloged events.

V. CONCLUSIONS

We have developed a new hierarchical Bayesian framework that is capable of recovering posterior probability distributions of compact-binary population hyper-parameters. These hyper-parameters encode details of stellar evolution, progenitor conditions, and the evolutionary paths taken to form systems that are detected by ground-based GW instruments such as Advanced LIGO and Advancdd Virgo.

Our methods fuse non-parametric (i.e. agnostic) modeling of GW parameter distributions with population synthesis simulations. Given a collection of population synthesis simulations of potential GW events, we first formed smoothed histograms of the binary parameters, stacked the vectors of histogram bin heights, then performed PCA to compress the bins into "features". This allowed significant dimensionality reduction while preserving the original distributions to high fidelity. We then trained GPs to interpolate the weights of these features across hyper-parameter space, so that we could emulate parameter distributions at any choice of population hyper-parameters between the simulated values. Using a GP allowed uncertainties in the interpolation training to be propagated through to subsequent statistical analyses. Other interpolant choices are possible; in future work we will explore the ability for a deep neural network to learn compact-binary distributions, and for such a network to be embedded in a population inference pipeline.

Having constructed a model for GW parameter distributions, we incorporated it into a hierarchical inference pipeline that used information from the distribution and rate of binary BH mergers in parameter space to discriminate compact-binary progenitor and evolutionary scenarios. We tested our pipeline on three case studies that successively increased in complexity and astrophysical realism. These ranged from a toy analytic model of binary component spin alignments, to publicly available population simulations, and finally to our own custom population synthesis simulations using a modified version of the publicly-available `BSE` code. In our final

study, we trained Gaussian processes on the 2-D distribution of binary BH chirp masses and redshifts across the hyper-parameter space of progenitor stellar metallicity, BH natal kicks, and common-envelope hardening efficiency. The recovered hyper-parameter posteriors were fully consistent with the injected values. We also performed a simple analysis on the existing Advanced-LIGO–Advanced-Virgo binary BH catalog, where we incorporated parameter measurement uncertainties to constrain progenitor metallicity to be $Z < 0.09\,Z_{\odot}$ at 90% credibility. (However, there are many caveats to this, and we quote it only to demonstrate the capabilities of our framework.)

The framework introduced here can be expanded and refined in many different ways. Further study is needed to understand how hyper-parameter measurement uncertainties will scale with the number of detected binaries, and how these compare with Fisher matrix approaches [115]. Furthermore, while we have carried out studies in controlled circumstances, full production-level analysis of real GW catalogs will require that several conditions be met: e.g. (*i*) the number of required training simulations should be determined through an iterative process, where GP uncertainties are investigated across hyper-parameter space to motivate new simulation locations; (*ii*) the number of binaries in each simulation should be large enough (ideally $\gtrsim 10^3$) to construct smoothed distributions that are representative of a large population. These refinements are important since we found that sampling the hyper-parameter space was challenging in large-event catalogs.

In this paper we mainly focused on binary BH systems, but our approach can be easily generalized to incorporate the relative observed fraction of BH-BH, NS-BH, and NS-NS systems as another means of discriminating evolutionary and progenitor conditions. Likewise, we only considered classical isolated binary evolution as the mechanism of compact-binary formation, but our framework could be applied to dynamical formation scenarios, allowing the details of many-body scattering in dense stellar clusters to be revealed. A mixture model would allow us to tease apart the sub-populations within a GW catalog that have evolved through each mechanism. With this method, the mixing fractions are just other hyper-parameters that can be estimated together with those describing the various channels. Unfortunately, the public version of `BSE` that we used does not provide information on component BH spins. We stress that inclusion of spins (and other parameters in general, like eccentricity) can be easily accommodated within our framework by carrying out informative training simulations.

We are entering a new source-rich era of GW astronomy, where catalogs of compact binary coalescences will reveal much about stellar astrophysics, including the processes underlying stellar evolution and the dynamics of dense stellar clusters. As third-generation ground-based detectors become a reality, so too will the opportunity to probe star formation rates across cosmic time, constrain cosmological parameters, understand the equation-of-state of nuclear matter, and use the huge event rates to limit modifications to GR. Furthermore, a space-based detector such as LISA will catalog hundreds of massive BH mergers, permitting reconstruction of massive BH seed formation scenarios and accretion efficiencies over cosmic time. Incorporating the detailed physics of population simulations into GW catalog analysis will allow for powerful statistical inference of the aforementioned processes. We hope that our framework lays the foundation for this exciting endeavor.

The code used to perform all analyses in this paper is publicly-available at github.com/stevertaylor/gw_catalog_mining, along with an example `jupyter` notebook for the toy model analysis in Sec. IV A.

ACKNOWLEDGMENTS

The authors thank Michele Vallisneri and Will Farr for useful discussions regarding Bayesian hierarchical modeling. We are grateful to Astrid Lamberts and Drew Clausen for providing us with a modified version of the BSE population synthesis code. S.R.T. acknowledges support from the NANOGrav project which receives support from NSF Physics Frontier Center award number 1430284. S.R.T thanks Erika Salomon for fruitful discussions. D.G. is supported by NASA through Einstein Postdoctoral Fellowship Grant No. PF6-170152 awarded by the Chandra X-ray Center, which is operated by the Smithsonian Astrophysical Observatory for NASA under Contract NAS8-03060. A majority of the computational work was performed on Caltech computer cluster "Wheeler" supported by the Sherman Fairchild Foundation and Caltech. Some of the computational work was performed on the Nemo cluster at UWM supported by NSF grant No. 0923409.

[1] B. P. Abbott *et al.* (LIGO and Virgo Collaboration), PRL **116**, 061102 (2016), arXiv:1602.03837 [gr-qc].
[2] B. P. Abbott *et al.* (LIGO and Virgo Collaboration), PRX **6**, 041015 (2016), arXiv:1606.04856 [gr-qc].
[3] B. P. Abbott *et al.* (LIGO and Virgo Collaboration), PRL **116**, 241103 (2016), arXiv:1606.04855 [gr-qc].
[4] B. P. Abbott *et al.* (LIGO and Virgo Collaboration), PRL **118**, 221101 (2017), arXiv:1706.01812 [gr-qc].
[5] B. P. Abbott *et al.* (LIGO and Virgo Collaboration), ApJ **851**, L35 (2017), arXiv:1711.05578 [astro-ph.HE].
[6] B. P. Abbott *et al.* (LIGO and Virgo Collaboration), PRL **119**, 141101 (2017), arXiv:1709.09660 [gr-qc].

[7] B. P. Abbott *et al.* (LIGO and Virgo Collaboration), PRL **119**, 161101 (2017), arXiv:1710.05832 [gr-qc].
[8] B. P. Abbott *et al.* (LIGO and Virgo Collaboration), ApJ **848**, L12 (2017), arXiv:1710.05833 [astro-ph.HE].
[9] B. Sathyaprakash *et al.*, CQG **29**, 124013 (2012), arXiv:1206.0331 [gr-qc].
[10] N. Yunes, K. Yagi, and F. Pretorius, PRD **94**, 084002 (2016), arXiv:1603.08955 [gr-qc].
[11] B. P. Abbott *et al.* (LIGO and Virgo Collaboration), PRL **116**, 221101 (2016), arXiv:1602.03841 [gr-qc].
[12] S. De, D. Finstad, J. M. Lattimer, D. A. Brown, E. Berger, and C. M. Biwer, (2018), arXiv:1804.08583 [astro-ph.HE].
[13] B. P. Abbott *et al.* (LIGO and Virgo Collaboration), (2018), arXiv:1805.11581 [gr-qc].
[14] B. P. Abbott *et al.* (LIGO and Virgo Collaboration), ApJ **848**, L13 (2017), arXiv:1710.05834 [astro-ph.HE].
[15] B. P. Abbott *et al.* (LIGO and Virgo Collaboration), Nature **551**, 85 (2017), arXiv:1710.05835.
[16] T. A. Apostolatos, C. Cutler, G. J. Sussman, and K. S. Thorne, PRD **49**, 6274 (1994).
[17] K. A. Postnov and L. R. Yungelson, LLR **17**, 3 (2014), arXiv:1403.4754 [astro-ph.HE].
[18] J. R. Hurley, C. A. Tout, and O. R. Pols, MNRAS **329**, 897 (2002), astro-ph/0201220.
[19] R. G. Izzard, C. A. Tout, A. I. Karakas, and O. R. Pols, MNRAS **350**, 407 (2004), astro-ph/0402403.
[20] K. Belczynski, V. Kalogera, F. A. Rasio, R. E. Taam, A. Zezas, T. Bulik, T. J. Maccarone, and N. Ivanova, ApJS **174**, 223 (2008), astro-ph/0511811.
[21] M. Spera, M. Mapelli, and A. Bressan, MNRAS **451**, 4086 (2015), arXiv:1505.05201 [astro-ph.SR].
[22] N. Giacobbo, M. Mapelli, and M. Spera, MNRAS **474**, 2959 (2018), arXiv:1711.03556 [astro-ph.SR].
[23] K. Breivik, K. Kremer, M. Bueno, S. L. Larson, S. Coughlin, and V. Kalogera, ApJ **854**, L1 (2018), arXiv:1710.08370 [astro-ph.SR].
[24] M. U. Kruckow, T. M. Tauris, N. Langer, M. Kramer, and R. G. Izzard, (2018), arXiv:1801.05433 [astro-ph.SR].
[25] S. Stevenson, A. Vigna-Gómez, I. Mandel, J. W. Barrett, C. J. Neijssel, D. Perkins, and S. E. de Mink, Nature Comm. **8**, 14906 (2017), arXiv:1704.01352 [astro-ph.HE].
[26] M. J. Benacquista and J. M. B. Downing, LLR **16**, 4 (2013), arXiv:1110.4423 [astro-ph.SR].
[27] S. R. Taylor, J. R. Gair, and I. Mandel, PRD **85**, 023535 (2012), arXiv:1108.5161 [gr-qc].
[28] X.-J. Zhu, E. Thrane, S. Osłowski, Y. Levin, and P. D. Lasky, (2017), arXiv:1711.09226 [astro-ph.HE].
[29] M. Zevin, C. Pankow, C. L. Rodriguez, L. Sampson, E. Chase, V. Kalogera, and F. A. Rasio, ApJ **846**, 82 (2017), arXiv:1704.07379 [astro-ph.HE].
[30] B. Farr, D. E. Holz, and W. M. Farr, ApJ **854**, L9 (2018), arXiv:1709.07896 [astro-ph.HE].
[31] D. Wysocki, J. Lange, and R. O'Shaughnessy, (2018), arXiv:1805.06442 [gr-qc].
[32] C. Talbot and E. Thrane, ApJ **856**, 173 (2018), arXiv:1801.02699 [astro-ph.HE].
[33] J. Roulet and M. Zaldarriaga, ArXiv e-prints (2018), arXiv:1806.10610 [astro-ph.HE].
[34] K. Belczynski, D. E. Holz, T. Bulik, and R. O'Shaughnessy, Nature **534**, 512 (2016), arXiv:1602.04531 [astro-ph.HE].
[35] B. P. Abbott *et al.* (LIGO and Virgo Collaboration), ApJ **818**, L22 (2016), arXiv:1602.03846 [astro-ph.HE].
[36] A. Lamberts, S. Garrison-Kimmel, D. R. Clausen, and P. F. Hopkins, MNRAS **463**, L31 (2016), arXiv:1605.08783 [astro-ph.HE].
[37] S. Stevenson, F. Ohme, and S. Fairhurst, ApJ **810**, 58 (2015), arXiv:1504.07802 [astro-ph.HE].
[38] D. Gerosa and E. Berti, PRD **95**, 124046 (2017), arXiv:1703.06223 [gr-qc].
[39] S. Stevenson, C. P. L. Berry, and I. Mandel, MNRAS **471**, 2801 (2017), arXiv:1703.06873 [astro-ph.HE].
[40] D. Wysocki, D. Gerosa, R. O'Shaughnessy, K. Belczynski, W. Gladysz, E. Berti, M. Kesden, and D. E. Holz, PRD **97**, 043014 (2018), arXiv:1709.01943 [astro-ph.HE].
[41] I. Mandel, W. M. Farr, A. Colonna, S. Stevenson, P. Tiňo, and J. Veitch, MNRAS **465**, 3254 (2017), arXiv:1608.08223 [astro-ph.HE].
[42] K. Heitmann, D. Higdon, C. Nakhleh, and S. Habib, ApJ **646**, L1 (2006), astro-ph/0606154.
[43] S. Habib, K. Heitmann, D. Higdon, C. Nakhleh, and B. Williams, PRD **76**, 083503 (2007), astro-ph/0702348.
[44] S. R. Taylor, J. Simon, and L. Sampson, PRL **118**, 181102 (2017), arXiv:1612.02817.
[45] Z Arzoumanian *et al.* (NANOGrav Collaboration), ApJ **859**, 47 (2018), arXiv:1801.02617 [astro-ph.HE].
[46] J. W. Barrett, I. Mandel, C. J. Neijssel, S. Stevenson, and A. Vigna-Gómez, in *Astroinformatics*, IAU Symposium, Vol. 325 (2017) pp. 46–50, arXiv:1704.03781 [astro-ph.HE].
[47] M. D. McKay, R. J. Beckman, and W. J. Conover, Technometrics **21**, 239 (1979).
[48] A. Singh, *pyDOE: The experimental design package for python* github.com/tisimst/pyDOE.
[49] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, 2006).
[50] D. J. MacKay, NATO ASI F **168**, 133 (1998).
[51] M. Ebden, (2015), arXiv:1505.02965 [math.ST].
[52] J. R. Gair and C. J. Moore, PRD **91**, 124062 (2015), arXiv:1504.02767 [gr-qc].
[53] C. J. Moore, C. P. L. Berry, A. J. K. Chua, and J. R. Gair, PRD **93**, 064001 (2016), arXiv:1509.04066 [gr-qc].
[54] R. O'Shaughnessy, PRD **88**, 084061 (2013), arXiv:1204.3117 [astro-ph.CO].
[55] R. van Haasteren and M. Vallisneri, PRD **90**, 104012 (2014), arXiv:1407.1838 [gr-qc].
[56] S. Ambikasaran, D. Foreman-Mackey, L. Greengard, D. W. Hogg, and M. O'Neil, IEEE Transactions on Pattern Analysis and Machine Intelligence **38** (2015), arXiv:1403.6015 [math.NA].
[57] T. Bayes, Philosophical Transactions of the Royal Society of London Series I **53**, 370 (1763).
[58] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, PASP **125**, 306 (2013), arXiv:1202.3665 [astro-ph.IM].
[59] D. Shoemaker *et al.*, dcc.ligo.org/LIGO-T0900288.
[60] B. P. Abbott *et al.* (LIGO and Virgo Collaboration), (2018), arXiv:1805.11579 [gr-qc].
[61] S. Vitale, D. Gerosa, C.-J. Haster, K. Chatziioannou, and A. Zimmerman, PRL **119**, 251103 (2017), arXiv:1707.04637 [gr-qc].
[62] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. J. Forteza, and A. Bohé, PRD **93**, 044007 (2016), arXiv:1508.07253 [gr-qc].

[63] K. K. Y. Ng, S. Vitale, A. Zimmerman, K. Chatziioannou, D. Gerosa, and C.-J. Haster, (2018), arXiv:1805.03046 [gr-qc].

[64] T. Dal Canton, A. H. Nitz, A. P. Lundgren, A. B. Nielsen, D. A. Brown, T. Dent, I. W. Harry, B. Krishnan, A. J. Miller, K. Wette, K. Wiesner, and J. L. Willis, PRD **90**, 082004 (2014), arXiv:1405.6731 [gr-qc].

[65] S. A. Usman, A. H. Nitz, I. W. Harry, C. M. Biwer, D. A. Brown, M. Cabero, C. D. Capano, T. Dal Canton, T. Dent, S. Fairhurst, M. S. Kehl, D. Keppel, B. Krishnan, A. Lenon, A. Lundgren, A. B. Nielsen, L. P. Pekowsky, H. P. Pfeiffer, P. R. Saulson, M. West, and J. L. Willis, CQG **33**, 215004 (2016), arXiv:1508.02357 [gr-qc].

[66] B. S. Sathyaprakash and B. F. Schutz, LLR **12**, 2 (2009), arXiv:0903.0338 [gr-qc].

[67] L. S. Finn and D. F. Chernoff, PRD **47**, 2198 (1993), gr-qc/9301003.

[68] L. S. Finn, PRD **53**, 2878 (1996), gr-qc/9601048.

[69] M. Dominik, E. Berti, R. O'Shaughnessy, I. Mandel, K. Belczynski, C. Fryer, D. E. Holz, T. Bulik, and F. Pannarale, ApJ **806**, 263 (2015), arXiv:1405.7016 [astro-ph.HE].

[70] J. Abadie *et al.* (LIGO and Virgo Collaboration), CQG **27**, 173001 (2010), arXiv:1003.2480 [astro-ph.HE].

[71] D. Gerosa, *gwdet: Detectability of gravitational-wave signals from compact binary coalescences* doi.org/10.5281/zenodo.889966.

[72] I. Mandel, PRD **81**, 084029 (2010), arXiv:0912.5531 [astro-ph.HE].

[73] D. W. Hogg, A. D. Myers, and J. Bovy, ApJ **725**, 2166 (2010), arXiv:1008.4146 [astro-ph.SR].

[74] M. R. Adams, N. J. Cornish, and T. B. Littenberg, PRD **86**, 124032 (2012), arXiv:1209.6286 [gr-qc].

[75] D. Foreman-Mackey, D. W. Hogg, and T. D. Morton, ApJ **795**, 64 (2014), arXiv:1406.3020 [astro-ph.EP].

[76] S. R. Taylor and J. R. Gair, PRD **86**, 023502 (2012), arXiv:1204.6739 [astro-ph.CO].

[77] M. Pitkin, C. Messenger, and X. Fan, ArXiv e-prints (2018), arXiv:1807.06726 [astro-ph.IM].

[78] I. Mandel, W. Farr, and J. R. Gair, ArXiv e-prints (2018), arXiv:1809.02063 [astro-ph.physics.data-an].

[79] W. Farr, I. Mandel, and J. R. Gair, *Selection effects* github.com/farr/SelectionExample.

[80] A. Sesana, J. Gair, E. Berti, and M. Volonteri, PRD **83**, 044036 (2011), arXiv:1011.5893 [astro-ph.CO].

[81] M. Fishbach, D. E. Holz, and W. M. Farr, ApJ **863**, L41 (2018).

[82] C. Talbot and E. Thrane, PRD **96**, 023012 (2017), arXiv:1704.08370 [astro-ph.HE].

[83] D. Gerosa, M. Kesden, E. Berti, R. O'Shaughnessy, and U. Sperhake, PRD **87**, 104028 (2013), arXiv:1302.4442 [gr-qc].

[84] D. Gerosa, R. O'Shaughnessy, M. Kesden, E. Berti, and U. Sperhake, PRD **89**, 124025 (2014), arXiv:1403.7147 [gr-qc].

[85] D. Trifirò, R. O'Shaughnessy, D. Gerosa, E. Berti, M. Kesden, T. Littenberg, and U. Sperhake, PRD **93**, 044071 (2016), arXiv:1507.05587 [gr-qc].

[86] C. L. Rodriguez, M. Zevin, C. Pankow, V. Kalogera, and F. A. Rasio, ApJ **832**, L2 (2016), arXiv:1609.05916 [astro-ph.HE].

[87] W. M. Farr, S. Stevenson, M. C. Miller, I. Mandel, B. Farr, and A. Vecchio, Nature **548**, 426 (2017), arXiv:1706.01385 [astro-ph.HE].

[88] M. Arca Sedda and M. Benacquista, ArXiv e-prints (2018), arXiv:1806.01285.

[89] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al., Physical Review Letters **116**, 241102 (2016), arXiv:1602.03840 [gr-qc].

[90] J. R. Hurley, O. R. Pols, and C. A. Tout, MNRAS **315**, 543 (2000), astro-ph/0001295.

[91] A. Lamberts, S. Garrison-Kimmel, P. Hopkins, E. Quataert, J. Bullock, C.-A. Faucher-Giguère, A. Wetzel, D. Keres, K. Drango, and R. Sanderson, (2018), arXiv:1801.03099.

[92] K. Belczynski, T. Bulik, C. L. Fryer, A. Ruiter, F. Valsecchi, J. S. Vink, and J. R. Hurley, ApJ **714**, 1217 (2010), arXiv:0904.2784 [astro-ph.SR].

[93] Planck Collaboration, P. A. R. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, J. G. Bartlett, and et al., A&A **594**, A13 (2016), arXiv:1502.01589.

[94] R. O'Shaughnessy, V. Kalogera, and K. Belczynski, ApJ **716**, 615 (2010), arXiv:0908.3635.

[95] K. Belczynski, D. E. Holz, C. L. Fryer, E. Berger, D. H. Hartmann, and B. O'Shea, ApJ **708**, 117 (2010), arXiv:0812.2470.

[96] Y. Chen, A. Bressan, L. Girardi, P. Marigo, X. Kong, and A. Lanza, MNRAS **452**, 1068 (2015), arXiv:1506.01681 [astro-ph.SR].

[97] C. L. Fryer, K. Belczynski, G. Wiktorowicz, M. Dominik, V. Kalogera, and D. E. Holz, ApJ **749**, 91 (2012), arXiv:1110.1726 [astro-ph.SR].

[98] H.-T. Janka, MNRAS **434**, 1355 (2013), arXiv:1306.0007 [astro-ph.SR].

[99] G. Hobbs, D. R. Lorimer, A. G. Lyne, and M. Kramer, MNRAS **360**, 974 (2005), astro-ph/0504584.

[100] S. Repetto, A. P. Igoshev, and G. Nelemans, MNRAS **467**, 298 (2017), arXiv:1701.01347 [astro-ph.HE].

[101] S. Repetto, M. B. Davies, and S. Sigurdsson, MNRAS **425**, 2799 (2012), arXiv:1203.3077 [astro-ph.GA].

[102] R. O'Shaughnessy, D. Gerosa, and D. Wysocki, PRL **119**, 011101 (2017), arXiv:1704.03879 [astro-ph.HE].

[103] C. L. Fryer, ApJ **522**, 413 (1999), astro-ph/9902315.

[104] C. L. Fryer and V. Kalogera, ApJ **554**, 548 (2001), astro-ph/9911312.

[105] B. Paczynski, in *Structure and Evolution of Close Binary Systems*, IAU Symposium, Vol. 73 (1976) p. 75.

[106] I. Iben, Jr. and M. Livio, PASP **105**, 1373 (1993).

[107] R. E. Taam and E. L. Sandquist, ARA&A **38**, 113 (2000).

[108] N. Ivanova, S. Justham, X. Chen, O. De Marco, C. L. Fryer, E. Gaburov, H. Ge, E. Glebbeek, Z. Han, X.-D. Li, G. Lu, T. Marsh, P. Podsiadlowski, A. Potter, N. Soker, R. Taam, T. M. Tauris, E. P. J. van den Heuvel, and R. F. Webbink, A&A Rev. **21**, 59 (2013), arXiv:1209.4302 [astro-ph.HE].

[109] X.-J. Xu and X.-D. Li, ApJ **716**, 114 (2010), arXiv:1004.4957 [astro-ph.SR].

[110] A. J. Loveridge, M. V. van der Sluys, and V. Kalogera, ApJ **743**, 49 (2011), arXiv:1009.5400 [astro-ph.SR].

[111] Z.-Y. Zuo and X.-D. Li, MNRAS **442**, 1980 (2014), arXiv:1405.4662 [astro-ph.HE].

[112] M. Dominik, K. Belczynski, C. Fryer, D. E. Holz, E. Berti, T. Bulik, I. Mandel, and R. O'Shaughnessy, ApJ **759**, 52 (2012), arXiv:1202.4901 [astro-ph.HE].

[113] K. Belczyński and T. Bulik, A&A **346**, 91 (1999), astro-ph/9901193.
[114] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization* (John Wiley and Sons, 2015).
[115] J. W. Barrett, S. M. Gaebel, C. J. Neijssel, A. Vigna-Gómez, S. Stevenson, C. P. L. Berry, W. M. Farr, and I. Mandel, MNRAS **477**, 4685 (2018), arXiv:1711.06287 [astro-ph.HE].