Deep learning to detect gravitational waves from binary close encounters: Fast parameter estimation using normalizing flows

Federico De Santi,[*] Massimiliano Razzano,[†] Francesco Fidecaro, Luca Muccillo, Lucia Papalini, and Barbara Patricelli

Dipartimento di Fisica "Enrico Fermi", Università di Pisa, I-56127 Pisa, Italy and Istituto Nazionale di Fisica Nucleare, Sezione di Pisa, I-56127 Pisa, Italy

(Received 29 January 2024; accepted 27 March 2024; published 7 May 2024)

A yet undetected class of gravitational wave signals is represented by the close encounters between compact objects in highly-eccentric ($e \sim 1$) orbits, that can occur in binary compact systems formed in dense environments such as globular clusters. The expected gravitational signals from these close encounters are short-duration pulses that would repeat over a much longer timescale in case of multiple passages at periastron. These sources represent a unique opportunity of exploring astrophysical formation channels as well as a different way of testing general relativity. Furthermore, in the case of binary systems containing neutron stars, the observation of these sources could help to constrain the neutron star equation of state, thanks to the signature left in the gravitational wave signal by the f-modes excitation that can occur during the encounter. The detection and parameter estimation of these signals is however challenging given the short duration of expected signals and the sensitivities of current ground-based gravitational wave interferometers. We present a novel approach to perform fast detection and parameter estimation of gravitational wave signals from binary close encounters that exploits probabilistic machine learning. We have used conditional normalizing flows to model complex probability distributions and therefore infer posterior distributions for the source parameters. This architecture is able to perform inference in a very short time and its output can be directly compared with classical methods. Fast detection and parameter estimation is very important as it could trigger electromagnetic follow-up campaigns and offer the possibility to study these events in a multimessenger context. To develop and test the algorithm, we have focused on the simulations of single bursts emission obtained using the Effective Fly-by formalism and embedded in the noise of Advanced LIGO and Virgo during their third observing run (O3). Our proposed model outperforms standard Bayesian methods in accuracy and is ~5 orders of magnitude faster, being able to produce 5×10^4 posterior samples in just 0.5 s. The results are extremely promising and constitute the first successful attempt for a fast and complete parameter estimation of binary close encounters using deep learning, offering a new approach to study the evolution of orbital parameters of compact binary systems.

DOI: 10.1103/PhysRevD.109.102004

I. INTRODUCTION

The detection of gravitational waves represents a revolution in the way we probe the Universe and provides a new and independent tool to investigate the physics of extreme compact objects. For instance, the first detection of gravitational waves from the coalescence of a binary black hole system, GW150914 [1], provided the observational proof of the existence of stellar-mass black holes with masses greater than $\simeq 25 M_\odot$ and established that binary black holes can form in nature and can merge within a Hubble time. Furthermore, the detection of gravitational waves from the event GW170817 and its associated electromagnetic counterparts marked the birth of a new era in multimessenger astrophysics [2,3]. The joint observation of electromagnetic and gravitational waves provided the first confirmation that binary neutron star coalescence are progenitors of short gamma-ray bursts [4], and allowed the investigation of the origin of heavy elements [5,6]. Furthermore, multimessenger observations of GW170817 offered a new way of investigating the equation of state of neutron stars [7,8], testing general relativity [9] and measuring the Hubble constant [10].

The third gravitational wave transient catalog (GWTC-3) [11] contains 90 events detected by Advanced LIGO and Virgo during the first three observing runs (O1, O2, O3) from 2015 to 2020. All these events are associated with the coalescence of compact binary systems (CBCs) containing black holes and/or neutron stars. More specifically, several dozens are consistent with binary black hole (BBH) systems. The growing population of BBHs observed

[*]f.desanti@studenti.unipi.it
[†]massimiliano.razzano@unipi.it

through gravitational waves allowed to perform population studies that seem to support the presence of more than one binary formation channel [12,13]. There seem to be two main formation channels [14]: BBHs can be the outcome of isolated binary evolution, i.e., they can form from the evolution of stars paired together at birth, or they can form dynamically, through strong stellar encounters in dense environments as young, globular, and nuclear clusters or active galactic nuclei. A deeper understanding of these different formation mechanisms is crucial in order to fully explain the BBH population so far observed.

Recent simulations of dynamical interactions in globular clusters have predicted the existence of populations of binaries merging with non-null eccentricity ($e > 0.05$) [15,16]. Despite gravitational wave emission being in general an efficient mechanism for the orbit circularization during the binary evolution, these works have revealed the existence of BBH subpopulations forming in orbits with eccentricities $e \sim 1$. We will refer to them as close encounters (CE).

Accurate measurement of the parameters of CE signals is of paramount importance to study dynamical formation channels as well as gravity in the strong field regime. At the moment no confident gravitational wave signal emitted during a CE has been detected, making these sources new and potentially interesting to search for [17,18].

Due to the high eccentricity of these systems, the expected gravitational wave emission differs from the chirplike waveform detected from CBCs. Eccentricity induces a modulation in the waveform that, in the limit $e \to 1$, transforms it into a series of repeated short duration bursts emitted during each periastron passage.

The burst-like nature of the signal, combined with the expected low signal-to-noise ratio, makes the detection of these sources particularly challenging. While current search strategies are based on unmodeled searches, Deep Learning has been proposed as a possible new approach to analyze these sources [19,20].

This paper present a novel approach to the detection and parameter estimation of gravitational waves from CEs based on probabilistic machine learning. Our approach exploits normalizing flows (NFs) to combine Bayesian inference methods with deep learning. This approach has been successfully tested on other types of sources. For instance, BBH coalescences have been studied with DINGO [21,22]. We will focus on single burst emission from encounters of binary black hole systems, as they are ones most likely to be detected by the current generation of interferometric detectors. We defer to subsequent work the application to the case of repeated bursts. The paper is organized in this way. In Sec. II we discuss the dynamical scenarios for the formation of CEs and their expected gravitational wave emission derived from the *effective fly-by* formalism. Section III introduces normalizing flows and their properties. In Sec. IV we discuss HYPERION, the NF-based pipeline that we have developed for parameter estimation using NFs. Section V contains the training on a simulated dataset and the resulting performance of the pipeline. Finally, Sec. VI discusses the results and limitations of this approach.

II. BINARY CLOSE ENCOUNTERS AS GRAVITATIONAL WAVE SOURCES

The canonical formation channel for BBH systems is via isolated binary evolution driven by stellar physics [23]. Stellar evolution further predicts the existence of a gap in the BH mass distribution from $50^{+20}_{-10} M_\odot$ to approximately $120 M_\odot$ because of pair-instability supernovae. The main uncertainties in the boundaries of this mass gap are related to limited knowledge of processes at play during the evolution of massive stars: e.g., the $^{12}\mathrm{C}(\alpha,\gamma)^{16}\mathrm{O}$ reaction [13].

However, population studies, made possible thanks to catalogs of observed gravitational wave events, have revealed a slightly different picture. In particular, the inferred distribution for the primary mass component in GWTC-3 does not exhibit a sharp drop at $\sim 50 M_\odot$ [13] as one would expect from the outlined formation channel. The presence of a tail at higher masses seems to suggest that a fraction of the observed BBHs could have formed through additional formation channels that have to be of dynamical origin, i.e., from N-body interaction between stars and/or black holes.

Besides the mass distribution, another ingredient that can provide clues to the formation channel is the spin orientation of the binaries. For instance, isolated field binary evolution is believed to produce components with preferably aligned spins [24] in contrast to dynamical encounters which can lead to isotropic spin-orbit misalignment [25]. There are currently evidences for the spin distribution to require misalignment as well as events with anti-aligned spins [13]; this could suggest that some of the observed BBHs formed dynamically, but further investigations are needed.

Therefore, this has led to the examination of these additional channels, which are possible in highly dense stellar environments. Examples of such environments are globular clusters which have central densities $\rho_c \geq 10^4 M_\odot\ \mathrm{pc}^{-3}$ [26], young stellar clusters with $\rho_c > 10^3 M_\odot\,\mathrm{pc}^{-3}$ [26,27], nuclear star clusters of galactic nuclei [28] as well as active galactic nuclei [29].

A. Dynamics in dense stellar environments

Given the high stellar density in globular clusters, single-single, binary-single and even binary-binary interactions can take place and influence the evolution of binary systems. These interactions have been studied through numerical N-body simulations and have revealed a wide spectrum of possible final states [30,31].

Recent simulations in globular clusters [15,31] have indeed confirmed that multiple resonant interactions can

lead to the formation of highly eccentric compact binaries. A subset of these binaries forms in a condition in which energy loss due to gravitational wave emission produces a capture: the inspiral phase of the binary speeds up, leading to a merger with a non-negligible eccentricity. More importantly, a subset of them is expected to merge within the LIGO-Virgo-KAGRA frequency range [16].

B. Highly eccentric compact binaries in globular clusters: Populations and rates

Dynamical interactions in globular clusters can produce different populations of merging systems, each one with its typical eccentricity. The dominant frequency $f_{\rm GW}^{\rm peak}$ at which these binaries emit gravitational waves is [16]

$$f_{\rm GW}^{\rm peak} = \frac{\sqrt{GM}}{\pi}\frac{(1+e)^{1.1954}}{[a(1-e^2)]^{3/2}} \tag{1}$$

with M being the total mass of the binary, a its semimajor axis, and e the eccentricity.

BBHs mergers formed through dynamical interactions in globular clusters fall into three major categories [16,31], depending on the timescale $T_{\rm GW}$ for gravitational wave emission to drive a binary to merge and the average timescale $T_{\rm SE}$ between two successive encounters. In particular they are defined as [16]:

$$T_{\rm GW} \propto a^4(1-e^2)^{7/2}, \qquad T_{\rm SE} \propto na^2\sigma\left(1+\frac{GM}{2a\sigma^2}\right) \tag{2}$$

where n and σ are the number density and the velocity dispersion in the cluster, respectively.

The first category of BBH mergers is that of ejected inspirals, which are binary systems that, by the recoil from close interaction, acquire a center of mass velocity that exceeds the escape velocity of the cluster and get ejected from it. These mergers produce gravitational waves with $f_{\rm GW}^{\rm peak} \leq 10^{-2}$ Hz while being characterized by a nonzero eccentricity ($e > 0.01$) [32,33]. For this reason, they are among the major sources detectable by LISA [34].

The in-cluster mergers are a second category of binaries merging inside the cluster due to dynamical encounters, but not due to significant emission of gravitational waves during the encounters. They can be of two kinds: 2-body and 3-body mergers. The former are binary black holes that survive a binary-single interaction with semimajor-axis and eccentricities such that their inspiral times are less than interaction times ($\sim 10^7$ years [35]). Their eccentricity is expected to be similar to that of ejected inspirals and the $f_{\rm GW}^{\rm peak}$ near the LISA sensitivity band [36]. The latter are still formed through binary-single interactions. However, their pericenter distance is perturbed in such a way that the energy lost over one orbit through gravitational wave radiation is larger than the initial energy of the 3-body system. Timescales associated with this process are thus much smaller ($\sim$1 year), which implies gravitational waves frequency peaks in the ground-based detector sensitivity bands [35,36].

Finally, the category of gravitational wave captures consist of binaries that inspiral and merge during a resonant interaction itself due to the strong emission of gravitational waves. This interaction can be a binary-single, binary-binary, or even a single-single. In the latter, two initially unbound objects experience an encounter on a hyperbolic orbit that causes the binary to become bound and rapidly merge. They typically result in $f_{\rm GW} \geq 10^{-1}$ Hz. However, this mechanism is also able to produce highly eccentric binaries ($e \sim 1$) that will merge within the sensitivity band of ground-based detectors with timescales $\mathcal{O}$(seconds). Given the high eccentricities of this last subset, some of them are close enough to the unbound limit to experience fly-by encounters [16]. The expected rate of eccentric BBH captures is expected to be 1–2 $\rm Gpc^{-3}\,yr^{-1}$ in the local universe ($z < 1$) [16].

C. Astrophysical relevance of CE observations

Close encounters carry distinctive signatures that can be used to differentiate between different formation channels, hence probing the underlying mechanisms responsible for the binary formation and merger. Tests of general relativity can also be carried out with such sources. For eccentric bound orbits, the smallest pericenter distance can be $r_p/M \sim 4$ ($G = c = 1$ units) corresponding to $v_p \sim 0.7c$ [37]. Therefore, CEs provide themselves as a unique laboratory to test general relativity in the strong-field regime: higher order effects such as radiation reaction and tides are indeed expected to become dominant. Other than that, eccentricity can be used to put constraints on alternative or modified theories of gravity [38]. Neutron star's equation of state can also be constrained if one of them is present in the binary. In the case of CBCs, the effects related to the equation of state become relevant only during the late inspiral and post-merger phase. In the case of eccentric inspiral, on the contrary, f-modes on the NS surface can be excited during each close interaction [39]. CE events could be potentially interesting also from a multimessenger point of view, either when neutron stars [40] and/or BHs are involved. As already mentioned, CEs can happen between two BHs embedded in the accretion disk of an active galactic nuclei [29] and, in such a gas rich environment, the merger can also yield a significant, detectable EM counterpart (see e.g. [41,42]).

CEs could also be the source of a stochastic background from primordial black holes. Close hyperbolic encounters from primordial black holes have been recently proposed as a detectable source for Einstein Telescope [43]. Being not resolvable, this emission results in overlapping bursts forming a stochastic background. In this work we will consider BBH gravitational wave captures at high

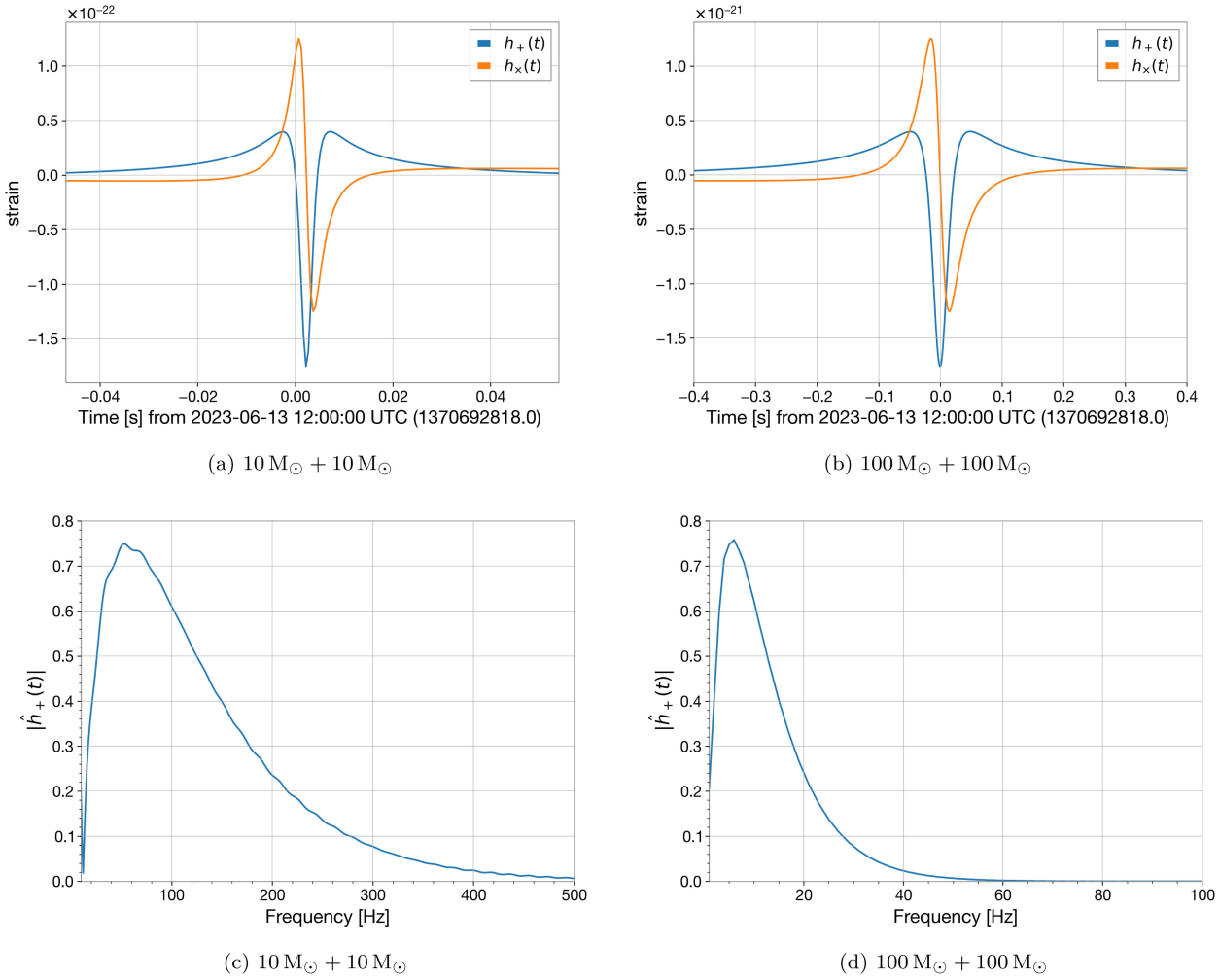
eccentricity ($e \sim 1$), since they are expected to be detectable with current ground-based interferometers.

D. Waveforms for eccentric close encounters

In order to be able to infer the parameters of a CE source, an accurate theoretical description of the gravitational wave signal emitted is needed. The presence of eccentricity, which is the defining feature of these system, poses several challenges. In first place, it makes mandatory to have accurate waveforms models. Indeed, even a small orbital eccentricity, if not correctly accounted for, is able to introduce systematic biases that exceed the statistical errors in parameter estimation [44]. As an example in [45] it has been shown that black hole captures might be misclassified as standard CBCs.

Currently, the most accurate gravitational wave waveforms are obtained through numerical relativity simulations, which have the drawback of being extremely computationally costly. This is due to the great velocities reached during the encounter which impose small integration steps. On the other hand, successive periastron passages happen on much wider timescales. Numerical relativity simulations available today, hence, only cover a limited number of orbits [46,47] and have shown that the gravitational wave emission consists in a series of repeated burst signals.

Since numerical relativity waveforms are too expensive to be exploited during an online analysis, it is crucial to also pursue an analytical approach. In order to account for relativistic effects such as radiation reaction, the post-Newtonian formalism is widely used. This method, which works well for binaries in quasicircular orbit, has difficulties in the high eccentricity limit since it is based on a post-circular expansion where the eccentric orbit is seen as a perturbation of a circular one. Previous attempts to describe eccentric waveforms in this way have been done in [48,49] up to eccentricities $\lesssim 0.8$ for widely separated binaries. However this approach suffers from post-Newtonian convergence issues when considering higher e or smaller



(a) $10\,\mathrm{M}_\odot + 10\,\mathrm{M}_\odot$

(b) $100\,\mathrm{M}_\odot + 100\,\mathrm{M}_\odot$

(c) $10\,\mathrm{M}_\odot + 10\,\mathrm{M}_\odot$

(d) $100\,\mathrm{M}_\odot + 100\,\mathrm{M}_\odot$

FIG. 1. Top: plus and cross polarizations waveforms obtained with the effective fly-by formalism. The captions indicate the BH masses, while other relevant parameters are $e = 0.9$ and $\bar{p} = 15$. Note the different timescales. Bottom: FFT of the plus polarizations above. We see that the signals lie in the LIGO and Virgo sensitivity band and that the increase of the total mass M results in a peak at lower frequencies.

separations [37]. Therefore the description of close encounter mergers is not fully feasible with it.

An alternative solution is represented by a formalism recently developed: the effective fly-by formalism [37,50]. The difference with respect to other analytical approaches is that the periastron passage in the eccentric orbit is obtained by perturbing a parabolic fly-by. That defines the post-parabolic approximation [50]. Hence, the effective fly-by formalism provides an accurate analytical description of the single burst emission at each periastron passage by modeling the single close passage as fly-by: i.e., a perturbation on a parabolic orbit. This method overcomes the issues of the post-circular approximation and is best suited for higher eccentricities. It is also possible to derive the whole inspiral waveforms from a single burst by adding many of these. In order to do so, it is necessary to include radiation reaction effects to track the evolution of the orbital parameters through time. Time-domain waveforms produced with this formalism (henceforth referred to as EFB-T) are given by [50]

$$h_{+,\times}(t) = -\frac{M^2\eta}{p[\ell(t)]d_L}\sum_{k=0}^{6}\sum_{n=0}^{2}\epsilon^n\Phi^{(n,k)}(\iota,\psi) + \mathcal{O}(\epsilon^3) \quad (3)$$

where M is the total mass of the binary, $\eta = m_1m_2/M^2$ the symmetric mass ratio, ι-ψ the inclination and polarization angles respectively, p is the semilatus rectum of the orbit which corresponds to the distance perpendicular to the semimajor axis to one of the focuses. In $G = c = 1$ units, it can be measured in $M_\odot$ units, and it is convenient to normalize it with respect to the total mass M: $\bar{p} \equiv p/M$. It is also related to the pericenter distance by

$$\bar{r}_p = \frac{\bar{p}}{1+e} \quad (4)$$

$\ell(t)$ is the mean anomaly defined as

$$\ell(t) = \frac{2\pi}{T_{\rm orb}(t)}(t - t_p) \quad (5)$$

with t_p the time of periastron passage and $T_{\rm orb}$ the orbital period. The relation $p[\ell(t)]$ accounts for radiation reaction effects at 2.5 post-Newtonian order (see Sec. III B in [50]). The waveforms so computed are valid only near t_p ($t \in [-t_{l=\pi}, t_{l=\pi}]$) and reproduce the parabolic limit as $e \to 1$.

Examples of the EFB-T plus and cross polarizations waveform are given in Fig. 1. From Eq. (3) the parameter that mainly affects both polarizations is the total mass M. With other parameters fixed, more massive binaries result in a longer and broader burst signal peaked at lower frequencies. Even so, the bursts have very short duration $\lesssim 1$ s, and an overall peak frequency in the range 10–100 Hz.

The good accuracy of these waveforms has been studied in [50] by comparing it with numerical waveforms at leading post-Newtonian order [51] and full numerical relativity.

III. NORMALIZING FLOWS FOR PARAMETER ESTIMATION

A. Basic definitions

The objective of Bayesian Inference in the context of gravitational wave data analysis is to obtain the posterior distribution for the parameters describing the signal. To compute it in the case of close encounters sources, we have exploited, in this work, the method of *normalizing flows* [52]. They are a powerful class of generative models capable of modeling complex probability distributions $p(\mathbf{x})$ out of simpler base distributions by means of a learned invertible transformation. The transformation can be conditioned on data thus making it possible to model surrogate posteriors $q(\boldsymbol{\theta}|\boldsymbol{s}) \approx p(\boldsymbol{\theta}|\boldsymbol{s})$. The key aspect of this approach is that it does not require any likelihood evaluation as the flow learns how to map $\boldsymbol{\theta}$ to the base distribution via a simulation-based process. Furthermore, inference requires only to evaluate the inverse transformation on samples from the base distribution, thus leading to a significant reduction in computational inference time.

To introduce the definition of a normalizing flow, let $\mathbf{x}$ be a vector in an input data space $\mathcal{X}$, distributed as $\mathbf{x} \sim p(\mathbf{x})$: a normalizing flow is then defined by an invertible map (bijection) $f_\phi\colon \mathcal{X} \to \mathcal{U}$ from the input data space $\mathcal{X}$ to a latent space $\mathcal{U}$ of a random variable $\mathbf{u} \sim \pi_\psi(\mathbf{u})$

$$\mathbf{x} \xrightarrow{f_\phi} \mathbf{u} \sim \pi_\psi(\mathbf{u}) \qquad \text{(forward pass)} \quad (6)$$

Our notation follows [53], with ϕ and ψ parameters f and π depend respectively upon. Since Eq. (6) is nothing but a change of variable, the probability distribution $p(\mathbf{x})$ can be expressed in terms of the base distribution as:

$$p(\mathbf{x}) = \pi_\psi(\mathbf{u})|\det \mathcal{J}_{f_\phi}| = \pi_\psi(f_\phi(\mathbf{x}))\left|\det\left(\frac{\partial f_\phi(\mathbf{x})}{\partial \mathbf{x}}\right)\right| \quad (7a)$$

$$\log p(\mathbf{x}) = \log \pi_\psi(f_\phi(\mathbf{x})) + \log\left|\det\left(\frac{\partial f_\phi(\mathbf{x})}{\partial \mathbf{x}}\right)\right| \quad (7b)$$

where $\mathcal{J}_{f_\phi} = (\frac{\partial f_\phi}{\partial \mathbf{x}})$ is the Jacobian of the transformation.

The map f_ϕ is learned by performing the forward pass specified by Eq. (6), then the *sampling* of $p(\mathbf{x})$ is straightforward and simply consists in evaluating the *inverse* f_ϕ^{-1} over samples from the base distribution

$$\mathbf{x} \xleftarrow{f_\phi^{-1}} \mathbf{u} \sim \pi_\psi(\mathbf{u}) \qquad \text{(inverse pass)} \quad (8)$$

This evaluation can be done as long as some conditions hold. First, $\pi_\psi(\mathbf{u})$ must be easy to sample and evaluate. To this scope, the uniform or Gaussian distribution are best

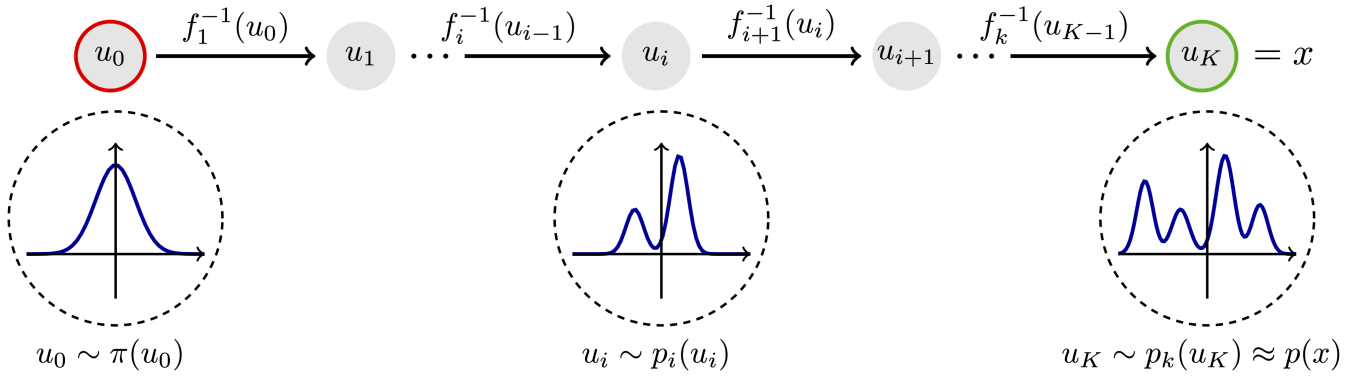


FIG. 2. A schematic representation of the inverse pass of a normalizing flow where the bijection is made up by a series of composite functions. During the inverse pass (sampling) the samples from the base distribution are gradually transformed in each step into a more complex distribution to match the target. Adapted from [54].

suited. Second, f_ϕ must be invertible, and third: f_ϕ and its inverse are differentiable. Furthermore, data and latent spaces share the same topology and dimensionality: the common choice is $\mathcal{X} = \mathcal{U} = \mathbb{R}^D$.

B. Expressive power and flexibility

It is interesting to consider whether a flow-based model can represent any distribution. If $p(\mathbf{x})$ and $\pi_\psi(\mathbf{u})$ are well behaved distributions satisfying the autoregressivity assumption:

$$p(\mathbf{x}) = \prod_{i=1}^{D} p(x_i|\mathbf{x}_{<i}), \quad p(x_i|\mathbf{x}_{<i}) > 0 \quad \forall\ i, \mathbf{x} \in \mathbb{R}^D \tag{9}$$

then there exists a diffeomorphism F that can map $\pi_\psi(\mathbf{x})$ into $p(\mathbf{x})$ [52]. Although this guarantees its existence, it does not provide a closed formula for F, so that it must be learned by optimizing a function f_ϕ. Therefore the expressive power of a normalizing flow, i.e. its ability to model complex distributions, strictly depends on the form of f_ϕ. Making flows more expressive can be achieved by increasing the flexibility of the bijection f_ϕ. For instance, given that a single function may not be sufficient, the whole bijection can be constructed as a combination of intermediate bijections:

$$f_\phi = f_{\phi_1}^{(1)} \circ f_{\phi_2}^{(2)} \circ \cdots \circ f_{\phi_K}^{(K)} \tag{10}$$

each one with its own set of parameters ϕ_i to be optimized. Under this assumption, the Jacobian can be factorized:

$$\mathcal{J}_{f_\phi} = \prod_{j=1}^{K} \mathcal{J}_{f_{\phi_j}^{(j)}} \tag{11}$$

Then Eq. (7b) reads

$$\log p(\mathbf{x}) = \log \pi_\psi(f_\phi(\mathbf{x})) + \sum_{j=1}^{K} \log |\det \mathcal{J}_{f_{\phi_j}^{(j)}}(\mathbf{u}_{j-1})| \tag{12}$$

This shows also the meaning of the name "normalizing flows": the input samples $\mathbf{x}$ undergoes a series of composite bijections to be gradually transformed into noise: i.e. $p(\mathbf{x})$ flows through each discrete step to be normalized. The reverse is true when computing the inverse to sample $p(\mathbf{x})$. Figure 2 gives a graphical representation of this concept. As will be discussed in Sec. III G, the bijections may be parametrized with the support of deep neural networks to increase expressiveness.

C. Likelihood-free inference

The main application of a normalizing flow model is probability density estimation and sampling, as stated by Eq. (7a). This approach is useful in cases where it is possible to have access to a collection of samples drawn from an unknown distribution that we would like to reconstruct. Indeed, by fitting the model through Eq. (6) then new samples can be generated as illustrated by Eq. (8). However the list of possible applications for such models does not end up here, as they can also perform variational inference. We will focus more on this kind of application as it fits our studying purposes.

Our goal is to infer probability distributions for a set of implicit parameters $\boldsymbol{\theta}$ that better describe some observation data $\mathbf{x}$. In our particular case $\mathbf{x} \equiv \mathbf{s}(t)$ is the strain time series containing the gravitational wave signal of a close encounter, and $\boldsymbol{\theta}$ the parameters of the physical system that generated it. From the Bayes theorem

$$p(\boldsymbol{\theta}|\mathbf{s}) \propto p(\boldsymbol{\theta}) p(\mathbf{s}|\boldsymbol{\theta}) \tag{13}$$

The posterior distribution $p(\boldsymbol{\theta}|\mathbf{s})$ is traditionally computed either with Monte Carlo Markov Chain (MCMC) or nested sampling by repeated evaluations of the likelihood $p(\mathbf{s}|\boldsymbol{\theta})$. This can become a bottleneck in many situations, either because the likelihood function can be costly to evaluate or because it may be not well defined thus preventing a tractable computation. Alternatively, normalizing flows provide themselves as a natural method to

approximate the posterior by producing a surrogate posterior $q(\boldsymbol{\theta}|\mathbf{s})$ in a tractable way. This can be done by making the bijection conditioned on the observed data.

$$p(\boldsymbol{\theta}|\mathbf{s}) \approx q(\boldsymbol{\theta}|\mathbf{s}) = \pi_\psi(f_\phi(\boldsymbol{\theta},\mathbf{s}))\left|\det\left(\frac{\partial f_\phi(\boldsymbol{\theta},\mathbf{s})}{\partial\boldsymbol{\theta}}\right)\right| \tag{14}$$

It is worth emphasizing that Eq. (14) does not require any likelihood evaluation to perform inference. The only requirement is to be able to simulate the data from a given set of parameters $\boldsymbol{\theta}^*$ extracted from a prior distribution:

$$\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta}), \qquad \mathbf{s}^* \sim p(\mathbf{s}|\boldsymbol{\theta}^*) \tag{15}$$

By this simulation process, the model indirectly incorporates both the prior over the parameters and the likelihood since the data points are generated accordingly. Therefore, this whole approach goes under the name of likelihood-free inference or even simulation-based inference [55]. As in other methods based on machine learning, inference is significantly faster since the computational cost is mostly during the training phase.

D. Training of normalizing flows

The training of normalizing flow models consists in optimizing the set of parameters $\boldsymbol{\phi}$ upon which the bijection f_ϕ depends by minimizing a suitable loss function. Given our purposes of inferring a surrogate gravitational wave posterior, in order for $q(\boldsymbol{\theta}|\mathbf{s}) \approx p(\boldsymbol{\theta}|\mathbf{s})$ it is necessary to minimize the distance between the two. The most straightforward measure of how close two distributions are is the Kullback-Leibler divergence [56]. The true posterior is in principle unknown but we can use the simulated set of samples $\{\boldsymbol{\theta}^{(i)},\mathbf{s}^{(i)}\}_{i=1}^N$ to minimize the *forward* KL divergence $\mathbb{KL}[p||q]$.

It is possible to derive an expression for the loss in the following way

$$\begin{aligned}
\mathcal{L} &= \mathbb{KL}[p(\boldsymbol{\theta}|\mathbf{s})||q_{\boldsymbol{\phi}}(\boldsymbol{\theta}|\mathbf{s})] \\
&= \int d\mathbf{s}\, p(\mathbf{s}) \int d\boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{s}) \log\left(\frac{p(\boldsymbol{\theta}|\mathbf{s})}{q_{\boldsymbol{\phi}}(\boldsymbol{\theta}|\mathbf{s})}\right) \\
&= \int d\mathbf{s}\, p(\mathbf{s}) \Bigg[\underbrace{-\int d\boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{s}) \log q_{\boldsymbol{\phi}}(\boldsymbol{\theta}|\mathbf{s})}_{\mathbb{H}[p||q_\phi]} + \underbrace{\int d\boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{s}) \log p(\boldsymbol{\theta}|\mathbf{s})}_{\mathbb{H}[p||p]=\text{cost}} \Bigg] \\
&\simeq -\int d\boldsymbol{\theta}\, p(\boldsymbol{\theta}) \int d\mathbf{s}\, p(\mathbf{s}|\boldsymbol{\theta}) \log q_{\boldsymbol{\phi}}(\boldsymbol{\theta}|\mathbf{s}) \\
&\simeq -\frac{1}{N}\sum_{i=1}^N \log q_{\boldsymbol{\phi}}(\boldsymbol{\theta}^{(i)}|\mathbf{s}^{(i)})
\end{aligned} \tag{16}$$

where $\mathbb{H}[p||q_\phi]$ is the differential cross-entropy between the two distributions and $\mathbb{H}[p||p]$ can be discarded, being constant with respect to the flow's parameters. At the fourth line we have applied the Bayes theorem to express the cross entropy in terms of the likelihood instead of the unknown posterior and in the last passage we leveraged the fact that we are in a simulation based context [cf. Eq. (15)] which implies that the integral can be approximated via Monte Carlo methods.

Therefore, minimizing the KL divergence is equivalent to minimizing the cross-entropy between p and q. By substituting Eq. (7b) we obtain the final formula

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^N \left[\log \pi_\psi(f_\phi(\boldsymbol{\theta}^{(i)};\mathbf{s}^{(i)})) + \log\left|\det\left(\frac{\partial f_\phi(\boldsymbol{\theta}^{(i)};\mathbf{s}^{(i)})}{\partial\boldsymbol{\theta}}\right)\right|\right] \tag{17}$$

So, minimizing the loss defined by Eq. (17) guarantees that the distribution inferred by the flow will converge to an optimal approximation for the true posterior. That relation remarks another time the likelihood-free nature of this approach since even to optimize the flow no likelihood evaluations are required: hence, the likelihood enters in the model via the simulated dataset.

Furthermore, the optimization of the parameters can be performed via stochastic gradient methods since unbiased estimators for the gradients are given by [52]:

$$\nabla_\phi \mathcal{L} \approx -\frac{1}{N}\sum_{i=1}^N \nabla_\phi \pi_\psi(f_\phi(\boldsymbol{\theta}^{(i)};\mathbf{s}^{(i)})) + \nabla_\phi \log|\mathcal{J}_{f_\phi}(\boldsymbol{\theta}^{(i)};\mathbf{s}^{(i)})| \tag{18a}$$

$$\nabla_\psi \mathcal{L} \approx -\frac{1}{N}\sum_{i=1}^N \nabla_\psi \pi_\psi(f_\phi(\boldsymbol{\theta}^{(i)};\mathbf{s}^{(i)})) \tag{18b}$$

Eq. (18b) is due to the fact that in some applications the base distribution can be learned together with the flow as

well. However in the case of likelihood-free inference is common practice to keep it fixed.

In deriving Eq. (17) we opted to minimize the forward KL divergence $\mathbb{KL}[p||q]$. In principle, there are other possible divergence measures that can be minimized: here, we motivate our choice. An alternative could have been the reverse KL divergence $\mathbb{KL}[q||p]$. This is typically adopted when the target density p is easy to evaluate but difficult to sample, which is not our case with posteriors over gravitational wave parameters. There is, however, a much more profound reason why the reverse is not the best option. First of all, KL divergence is *not* symmetrical. Thus, minimizing either one or the other leads to different results, as the optimized distribution will show different behaviors. More specifically, the forward KL is *mass covering* while the reverse is *mode seeking*. An intuitive explanation can be suggested. In the forward case, in order for KL to not diverge, $q > 0$ whenever $p > 0$, meaning it must cover the whole support of p. Conversely, in the reverse case, being p at the denominator: $q = 0$ whenever $p = 0$, thus forcing q to seek for the dominant mode in p. In the case of a multimodal distribution, as gravitational wave posteriors are, a mass covering approximant is preferable since it will not exclude less dominant modes that could provide interesting information.

E. Normalizing flows for gravitational wave data analysis

Two main algorithms are currently exploited to infer the Bayesian posteriors over gravitational wave parameters: MCMC and nested sampling. Both are based on Markov Chains and obtain samples from $p(\boldsymbol{\theta}|\mathbf{s})$ by means of repeated likelihood evaluations. This implies several computational drawbacks. First of all, the computational efficiency of these algorithms is severely limited by waveform generation, which can take about $10^{-3}\ \mathrm{s} \lesssim \langle\tau\rangle \lesssim 1\ \mathrm{s}$ [57], depending on the particular waveform model used. This, combined with the elevated number of required likelihood evaluations $\mathcal{O}(10^7)$ [58], gives a hint about the amount of time required to perform an analysis. Second, being based on Markov Chains, the produced samples show correlation, which has to be accounted for, thus reducing the number of effective samples. The high inference time is perhaps the most relevant limitation since it also impacts multimessenger observations as an early warning strategy is hardly implementable. Furthermore, the typically adopted Gaussian likelihood (see, e.g., [59]) assumes Gaussian (wide sense) stationary noise in the detector. Such a condition is not always completely satisfied as detectors may manifest both non-Gaussianities and nonstationarities like the frequent short transients known as glitches. Therefore, if the noise assumptions are violated, the whole analysis can be affected by biases. Parameter estimation analyses typically require a precise knowledge of the waveform models. In the case of close encounters, where uncertainties exist on the waveform modeling, it has been shown that the recovery of parameters (e.g., the masses) is limited by a small number of accessible bursts during the inspiral [60]. A NF-based approach can leverage the generalization capabilities of deep neural networks to better recover the parameters with a limited amount of information, providing, at the same time, reduced inference times.

Finally, computational efficiency will become a key aspect of data analysis in future observing runs as well as in the third-generation detector era. As a consequence of the higher sensitivity of future instruments, it is expected a $\sim 10^3$ increase in the event rate $\mathcal{R}$. As an example, $\mathcal{R} \gtrsim 10^5$ events/year for the Einstein Telescope [61]. Faster and more efficient algorithms will be crucial for the success of those experiments.

Normalizing flows provide themselves as a valid alternative able to supply to the limitations of traditional methods. In fact, as we discussed in Sec. III C, the cost of inference is completely amortized as likelihood evaluations are not required, and expensive waveform computations are performed only once during training. The fact of being a simulation-based inference has another implication worth emphasizing: it does not suffer from the limiting assumption about Gaussianity and stationarity of the noise, provided that an adequate description is available.

F. Model selection with normalizing flows

Another kind of analysis that strictly depends on parameter estimation is model selection (or hypothesis testing), which in the case of gravitational waves may refer to signal detection, i.e., testing the hypothesis of the presence or absence of a signal in the strain, or even discriminating between two waveform models what is better at describing the data. This is done by computing the Bayes factor $\mathcal{B}_{12} = \mathcal{Z}_1/\mathcal{Z}_2$ which compares the evidences (or marginal likelihoods) of the two hypothesis. Furthermore, when computed in the case of the null hypothesis of having only noise ($\mathcal{Z}_2 = \mathcal{L}_{\mathrm{noise}}$), $\mathcal{B}_{12}$ can be exploited as a detection statistic.

Although the product of a normalizing flow model is a direct approximation of the posterior, the evidence can be estimated as well through importance sampling, which is nothing but a Monte Carlo estimate. More precisely

$$\mathcal{Z} = \int d\boldsymbol{\theta} p(\boldsymbol{\theta}) p(\mathbf{s}|\boldsymbol{\theta}) = \int d\boldsymbol{\theta} \frac{p(\boldsymbol{\theta}) p(\mathbf{s}|\boldsymbol{\theta})}{q(\boldsymbol{\theta}|\mathbf{s})} q(\boldsymbol{\theta}|\mathbf{s}) \tag{19}$$

By sampling the flow posterior $q(\boldsymbol{\theta}|\mathbf{s})$, which is optimized by minimizing the mass covering forward KL divergence, we can get an estimator of the evidence from importance sampling weights.

$$\hat{\mathcal{Z}} = \frac{1}{N} \sum_{i=1}^{N} \frac{p(\boldsymbol{\theta}_i) p(\mathbf{s}|\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i|\mathbf{s})} = \frac{1}{N} \sum_{i=1}^{N} w_i \tag{20}$$

The only disadvantage is that w_i relies on the analytical likelihood to be computed. However, since they can be computed separately, the whole procedure can be parallelized in principle, reducing its computational cost.

G. Constructing the flow

We now discuss how normalizing flows can be constructed by implementing the bijection f_ϕ to be expressive and computationally efficient at the same time. When referring to computational efficiency, the interest is to find a function whose Jacobian, actually its determinant, is fast to compute. The function must also be easy to invert and rapid to evaluate both in the forward and inverse pass. On the other hand expressiveness refers to a sufficiently flexible transformation able to deal with highly complex distributions. More in general since $f_\phi\colon \mathbb{R}^D \to \mathbb{R}^D$ acts on D-dimensional vectors, it has the general form:

$$u_i = g_{\phi_i}(\theta_i; \mathbf{\Theta}_i), \qquad \mathbf{\Theta}_i = c_i(\boldsymbol{\theta}) \tag{21}$$

where c_i is the *conditioner*, which specifies how the bijection acts on the various dimensions and, in particular, on which set of components $\mathbf{\Theta}_i$ does θ_i depends. It is not required for it to be a bijection. g_ϕ is instead the *transformer*: a monotonic function, hence invertible, that actually transforms the input variables. The set of parameters $\boldsymbol{\phi}$ of f_ϕ contains both parameters of the conditioner and transformer. Since, however, the conditioner is typically specified before the training and it is not part of the optimization, its are just hyperparameters. For this reason, henceforth, we'll refer to $\boldsymbol{\phi}$ as the parameters of the transformer only.

The most simple flow that can be constructed is the so-called *element-wise flow* whose conditioner treats each vector dimensions independently Eq. (22).

$$\begin{cases} u_1 = g_{\phi_1}(\theta_1) \\ u_2 = g_{\phi_2}(\theta_2) \\ \quad\vdots \\ u_D = g_{\phi_D}(\theta_D) \end{cases}, \qquad \det \mathcal{J}_{f_\phi} = \prod_{i=1}^{D} \frac{\partial g_{\phi_i}}{\partial u_i} \tag{22}$$

This flow is efficient both in the forward and inverse pass due to the simple Jacobian: being a diagonal matrix, its determinant is just the product of the diagonal. However, it lacks expressiveness since each component is transformed independently. Hence, it will not be able to capture all the eventual dependencies and degeneracies among the various elements. In the case of gravitational waves, there are a lot of dependencies between parameters. As an example, recall from Eq. (3) that in the case of close encounters, the strain amplitude is $h_{+,\times}(t) \propto M^2/d_L$ which induces a degeneracy between the total mass and the luminosity distance. Other degeneracies can arise, for instance, when considering the localization of the source and the antenna pattern response of the detectors. There are other architectures able to deal with such situations, like the *autoregressive* conditioner or the *coupling layers*.

The former, in particular, models the dependencies between variables assuming an autoregressive structure Eq. (9) where each component θ_i depends upon $\theta_{j<i}$ components. With this assumption the bijection Eq. (21) becomes

$$u_i = g_{\phi_i}(\theta_i) \quad \text{with} \quad \phi_i = F(\boldsymbol{\theta}_{1:i-1}) \tag{23}$$

In most cases, g_ϕ is taken to be an analytical invertible function whose parameters ϕ are the output of a neural network here denoted with F [62].

The autoregressive transformation Eq. (23) is characterized by having a low triangular Jacobian

$$\mathcal{J}_{f_\phi}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial \mathrm{u}_1}{\partial \theta_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{A} & & \frac{\partial \mathrm{u}_D}{\partial \theta_D} \end{bmatrix} \tag{24}$$

hence making the computation of its determinant equivalent to the element-wise flow Eq. (22).

Nevertheless, the whole architecture of autoregressive flows manifests inefficiency when computing the inverse transformation (inference) as it takes a *recursive* structure. In fact, the sampling of θ_i from $\mathbf{u}$ requires to have already sampled $\boldsymbol{\theta}_{1:i-1}$ thus turning this operation in a sequential and non parallelizable one: see e.g. Fig. 3 in [52]. The computational cost scales in particular as $\mathcal{O}(D)$. It is, therefore, an unavoidable aspect of autoregressive flow to have either one of the two passes to be inefficient.[1] Although they are, in principle, the most expressive since they are able to account for any dependence in the variables, the computational cost either for training or sampling scales badly with high dimensional inputs.

Coupling layers were introduced in [53] to overcome the efficiency limitations of autoregressive flows while maintaining their expressiveness. The idea behind a coupling layer is to split the parameter space in two equally dimensional subsets $\boldsymbol{\theta} = (\boldsymbol{\theta}^d, \boldsymbol{\theta}^{D-d})$ with $d \simeq D/2$. The second half is then transformed element-wise and conditioned on the first half, which is mapped through an identity.

$$\begin{cases} \mathbf{u}_{1:d} = \boldsymbol{\theta}_{1:d} \\ \mathbf{u}_{d+1:D} = g_\phi(\theta_{d+1:D}; \boldsymbol{\theta}_{1:d}) \end{cases} \tag{25}$$

[1]It has been proposed indeed a slight variation of this flow which is the inverse autoregressive flow [63]. The recursive structure is moved from the inverse to the forward pass, but it cannot be removed.

The Jacobian is still a low triangular matrix

$$\mathcal{J}_{f_\phi}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbb{I}_d & \mathbf{0} \\ \mathbf{A} & \mathbf{D}_{D-d} \end{bmatrix}, \qquad \mathbf{D} = \mathrm{diag}\left[\frac{\partial \mathrm{u}_{d+1:D}}{\partial \theta_{d+1:D}}\right] \tag{26}$$

However, since the upper left block is simply the identity matrix, the computational cost scales as $\mathcal{O}(D-d)$. It turns out that computing both forward and inverse is an efficient operation that can be further parallelized. The only drawback of this layer is that a single one is not sufficient, as only half of the components get actually transformed. To enhance the expressiveness, it is possible to stack multiple of these layers with random permutations of $\boldsymbol{\theta}$ indexes in between. Hence, f_ϕ is given by Eq. (10) and the full Jacobian by Eq. (11). If the number K of layers is sufficient, the output will be equivalent to an autoregressive one due to the fact that, in the end, each component is transformed, being conditioned on every other component. As a "rule of thumb," K should at least be equal to D.

Coupling layers provide themselves as the optimal choice both in terms of expressiveness, flexibility, and computational cost: in fact, both training and sampling are equally fast. Moreover, they are also relatively easy to implement.

We now describe the invertible transformation. Any strictly monotonic function, being invertible, can be applied, provided, however, it is differentiable and with an easy-to-compute inverse. In the continuation of this discussion, we will consider two of the most widely adopted. *Affine transformations* were among the first functions to be proposed as suitable transformers. The same work introducing coupling layers adopted this form exploiting exponential rescaling [53,64]:

$$\begin{cases} \mathbf{u} = g_\phi(\boldsymbol{\theta}) = \boldsymbol{\theta} \odot \exp\left[s(\boldsymbol{\theta})\right] + t(\boldsymbol{\theta}) \\ \boldsymbol{\theta} = g_\phi^{-1}(\mathbf{u}) = \left[\mathbf{u} - t(\mathbf{u})\right] \odot \exp\left[-s(\mathbf{u})\right] \end{cases} \tag{27}$$

In Eq. (27), $\odot$ denotes the Hadamard (or element-wise) product. The parameters of this function are hence $\phi = \{t, s\}$: shift and scale.

Combined with coupling layers, the transformation of Eq. (27) has proven to be flexible and expressive enough to model complex distributions as images [65] or even audio waveforms [66].

Another flexible transformation was introduced in [67] as an avenue to model extremely complex and multimodal distributions while retaining the property of being analytical, differentiable, and easy to invert. The idea is to map an interval $[-B, B] \subset \mathcal{X}$ into $[-B, B] \subset \mathcal{U}$ by interpolating a rational quadratic spline between a set of sorted knots $\{x_k, y_k\}_{k=0}^{K}$ where both the knots and their internal derivatives $\{\delta_k\}_{k=1}^{K-1}$ are parametrized as the output of a neural network. Computing the inverse requires solving a 2nd order equation, which can be done analytically (see Eqs. (6)–(8) in [67]). This kind of transformation is extremely flexible, and it naturally induces multimodality by increasing the number K of knots. Therefore, it has been mainly applied in the context of image generation.

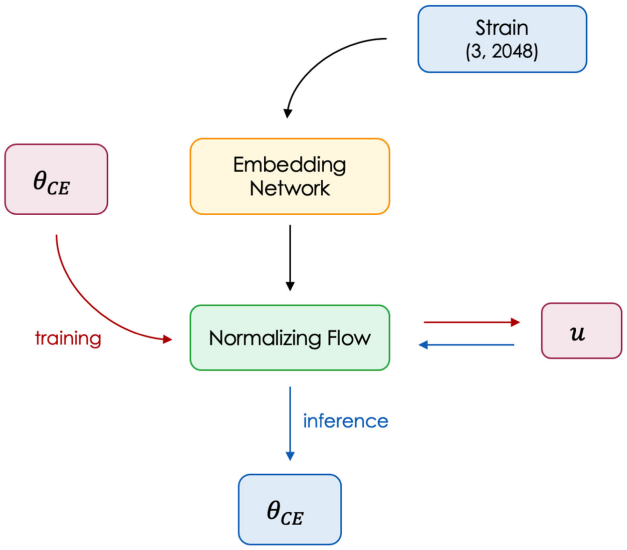


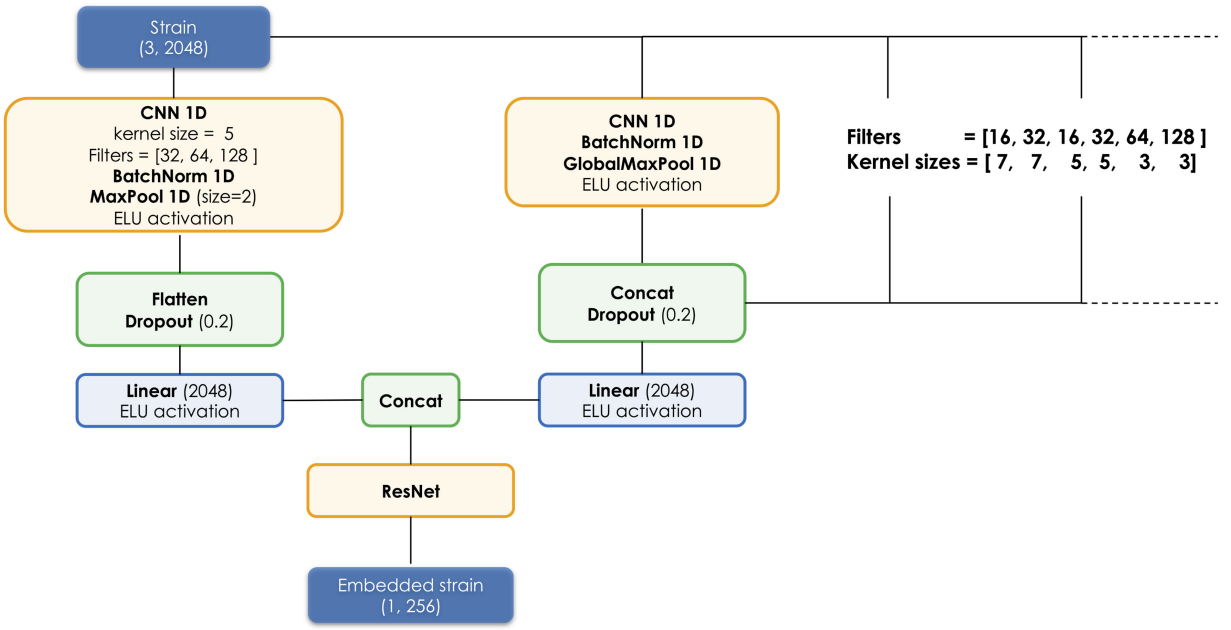
FIG. 3. Schematic overview of HYPERION which is composed of a normalizing flow and embedding neural network acting on input strain data. Solid arrows represents input-ouput relations: red apply during training, blue ones when performing inference while black ones are always present.

IV. HYPERION's ARCHITECTURE

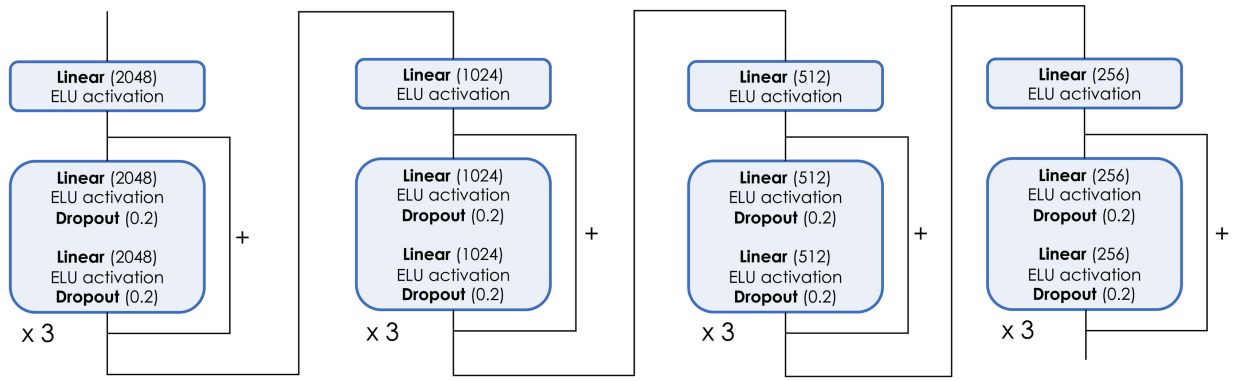
We present here the "*hyperfast close encounter inference from observations with normalizing-flows*" pipeline (HYPERION). This pipeline takes as input 1 s of whitened strain time series and returns as output samples from the posterior probability $p(\boldsymbol{\theta}_{\mathrm{CE}}|\boldsymbol{s})$. More specifically, the parameters over which it makes inference are: the total mass M, mass ratio q, eccentricity e_0,[2] semilatus rectum $\bar{p}_0$, luminosity distance d_L, the time of periastron passage δt_p, right ascension α and declination δ. The general structure of HYPERION is depicted in Fig. 3 along with input/output relations between its building blocks.

The core of the model is a normalizing flow, which reconstructs the posterior distribution. Given that it is a conditional probability distribution, the flow must be supplied with the most informative context as possible. Therefore, we introduced in the model another building block, fundamental as well: an embedding neural network. Acting as a feature extractor, its primary task is to extrapolate the information in the noisy strain time series and to compress it to a lower dimensional form. This procedure has the purpose of filtering out all the irrelevant features, mainly the noise content. Other than the embedding one, other deep

[2]the subscript 0 refers to the value when the mean anomaly $\ell = 0$, i.e., the periastron passage.



(a)



(b)

FIG. 4. (a): The general architecture of the embedding network is composed of two CNN blocks acting in different ways and a ResNet block that efficiently compresses the extracted features into a (1,256) dimensional tensor. More specifically, the CNN block on the left extracts features related to the signal morphology, while the other on the right focuses more on temporal correlated patterns. (b): Detailed architecture of the ResNet block.

neural networks are implemented in the normalizing flow itself, thus making our model reach the number of ∼180 millions of trainable parameters. HYPERION was developed with PYTHON 3.10 and PYTORCH 2.1.0 [68].

A. The embedding network

The presence of such an element in our model can be justified by the following reason. In the process of likelihood-free inference, the likelihood enters indirectly as the result of a simulation procedure. It means that the NF is able to determine the best mapping $f_\phi: \Theta \to \mathcal{U}$ based on the similarity between the joint samples $\{\boldsymbol{\theta}_{\mathrm{CE}}^{(i)}, \mathbf{s}^{(i)}\}$ that it is supplied with. A raw data representation, like the strain time series, is not the optimal choice, even if whitened. That is both because of the low signal to noise ratio for CE signals and because of the morphology of the signal itself, which does not show directly a clear dependency on all the $\boldsymbol{\theta}_{\mathrm{CE}}$ parameters. Hence, a feature extractor is necessary. The overall architecture of the embedding network shown in Fig. 4(a) is the result of several optimizations and improvements.

It is composed of two convolutional neural network (CNN) blocks that perform the feature extraction from the input time series: 1 s sampled at $f_s = 2048$ Hz with each of the 3 channels corresponding to a given interferometer

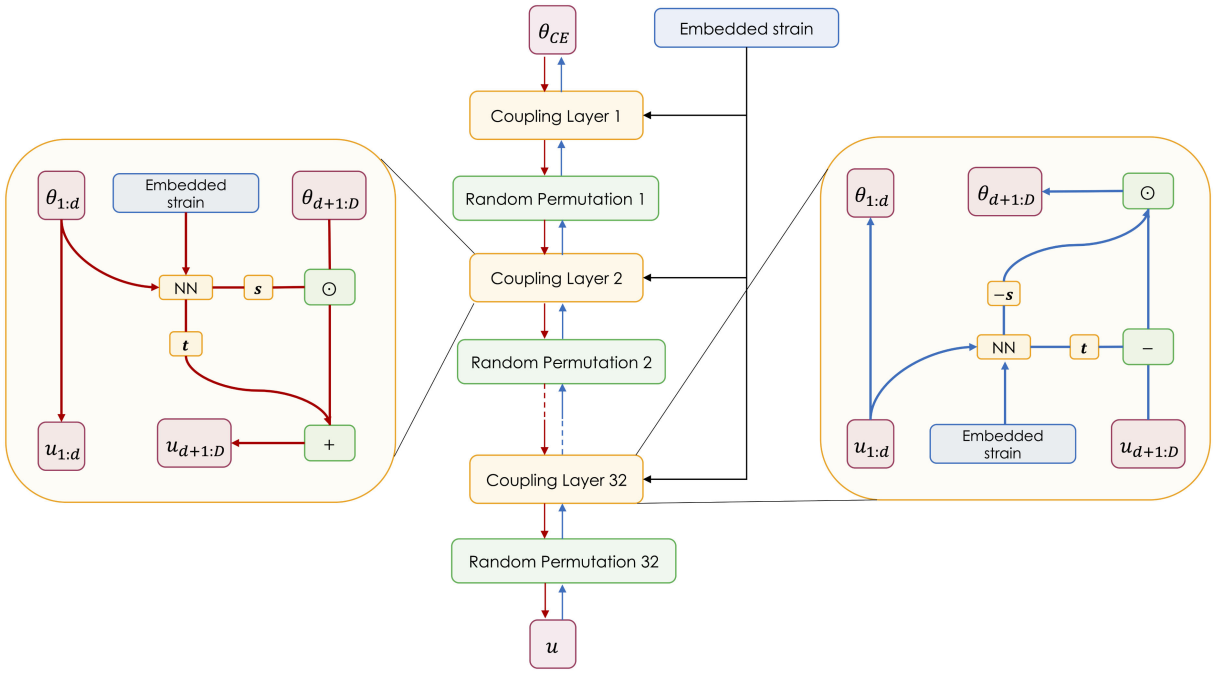


FIG. 5. The architecture of the normalizing flow implemented in HYPERION which consists of a stacking of 32 coupling layers with affine transformations. The red arrows refer to training, while the blue ones to inference (inverse of the transformation).

(H1, L1, V1). The first block consists of three 1D convolutional layers with fixed kernel size $= 5$ and [32, 64, 128] numbers of filters respectively. Between each layer, there are pooling layers and also batch normalization layers, whose addition was found to be beneficial. This first block is able to learn features related to the shape or morphology of the signal which are relevant for a subset of $\boldsymbol{\theta}_{\mathrm{CE}}$. Given, however, the translational invariance of the neuron's response in such block, it is unable to learn features that are mostly correlated with their.

Indeed, in our earlier experiments, the inference about $\{\delta t_p, \alpha, \delta\}$ was not great as we were simply recovering their priors. The reason for the effect on sky localization can be easily understood by the fact that it is determined by the relative shifts in arrival time of the signal in each detector. For long signals, like those produced by coalescences, multiple time instants can be compared: i.e., the information spreads over a wide temporal interval. For a burst signal, on the contrary, all the emission is concentrated over a small time interval, which implies that the sky localization information strongly correlates with δt_p itself.

This motivated introducing a parallel CNN block acting differently from the former. The crucial difference is that many convolutional layers, with different filters and kernel sizes, slide independently over the time series and after each of them a global max pooling layer keeps the maximum neuron's response. The resulting outputs are then concatenated together, and in this way, the temporal information about the neuron's response gets preserved.

The output of each CNN block is then passed to linear layers with 2048 neurons that are subsequently concatenated and then compressed into a final layer with 256 output neurons by means of the ResNet block [Fig. 4(b)]. This block is composed of four sub-blocks sized [2048, 1024, 512, 256] respectively, each one containing 3 skip connections. In contrast to regular linear layers, skip connections proved to be more efficient at compressing the dimensionality of the network's output without loss of meaningful information. Regarding activation functions, we have found the best results with the ELU rather than ReLU. To reduce at minimum the chances of overfitting the embedding network, especially the ResNet block, makes extensive use of dropout layers.

We decided to exploit CNNs and not recurrent neural networks (RNNs) mainly for two reasons. RNNs are suited for the analysis of long temporal correlated sequences. In our case, as already explained, the information is localized in time. CNNs are better suited to extract feature on different timescales, therefore they have been proposed as a viable machine learning method for gravitational wave data analysis. Furthermore, the recurrent structure of RNNs negatively impacts the computational cost of inference.

B. The normalizing flow

The NF implemented in HYPERION adopts Coupling Layers, given their properties and computational efficiency, combined with affine transformations Eq. (27). The whole normalizing flow scheme is shown in Fig. 5 where we made explicit the structure of the affine coupling layer Eq. (28).

$$\begin{aligned}(\text{training}):&\begin{cases}u_{1:d}=\theta_{1:d}\\ u_{d+1:D}=\theta_{d+1:D}\odot\exp\left[s_{d+1:D}(\theta_{1:d})\right]+t_{d+1:D}(\theta_{1:d})\end{cases}\\ (\text{inference}):&\begin{cases}\theta_{1:d}=u_{1:d}\\ \theta_{d+1:D}=[u_{d+1:D}-t_{d+1:D}(u_{1:d})]\odot\exp[-s_{d+1:D}(\theta_{1:d})]\end{cases}\end{aligned}\tag{28}$$

The architecture consists of 32 layers: a sufficiently high number to guarantee a proper mixing between all the $\boldsymbol{\theta}_{\mathrm{CE}}$ components in order to capture all the dependencies and degeneracies. In between every coupling layer, a random permutation shuffles the parameter's space indexes. This can be seen as an additional transformation with a Jacobian equal to $\mathbb{1}$. The permutation matrices are then saved as parameters of the model for reproducibility. We also tested rational quadratic splines, given their expected expressiveness, but they turned out to be sub-optimal. The posteriors produced were excessively multimodal, with clear signs of either underfitting or overfitting in some cases: this was also confirmed by the training and validation losses during the optimization.

The affine couplings depend upon two parameters: scale $\boldsymbol{s}$ and shift $\boldsymbol{t}$. Both of them are the output of a fully connected neural network (Fig. 6), which takes as input both the identity-mapped parameters and the embedded strain. In our implementation, each layer has its own network that is optimized independently for an overall more precise inference. The network for the scale and shift parameters are nearly identical except for the activation function. While the shift's one adopts the ELU, the other one adopts the tanh to prevent numerical instabilities that can arise otherwise due to the fact that $\boldsymbol{s}$ enters into an exponential. The hyperbolic tangent is also a better choice than the Sigmoid since it allows both ≤ 1 and ≥ 1 scale factor values.

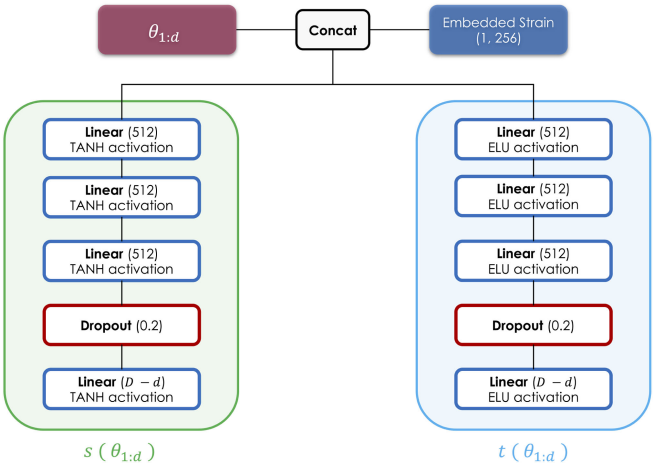


FIG. 6. Neural network architecture for the affine coupling layer Eq. (28). Each of the 32 coupling layers in HYPERION contains such a network that gets optimized independently. Note the different activation functions for the scale parameter branch.

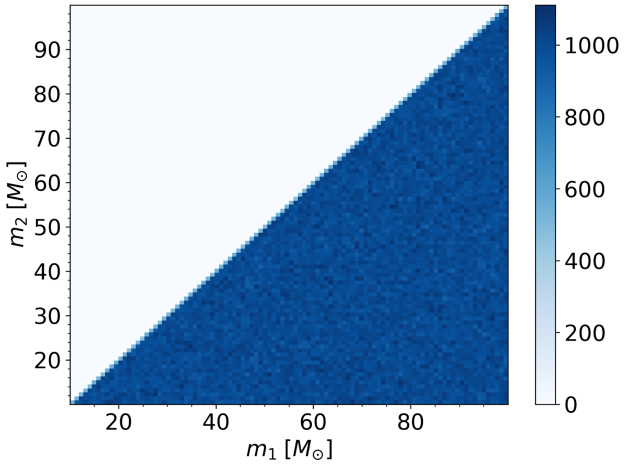
C. Pre- and postsamples processing

Since the various parameters in $\boldsymbol{\theta}_{\mathrm{CE}}$ might have wide and different numerical ranges, a direct usage of their strict value would certainly result in numerical instabilities when fed to the neural networks. For that reason, each parameter is rescaled to have zero mean and unit variance. This reduces their numerical range while keeping intact the shape of their prior distribution at the same time. Means $\boldsymbol{\mu}$ and standard deviations $\boldsymbol{\sigma}$ are computed from the training dataset and saved as model hyperparameters. At the end of the inference phase, all the samples are brought back to their original physical range.

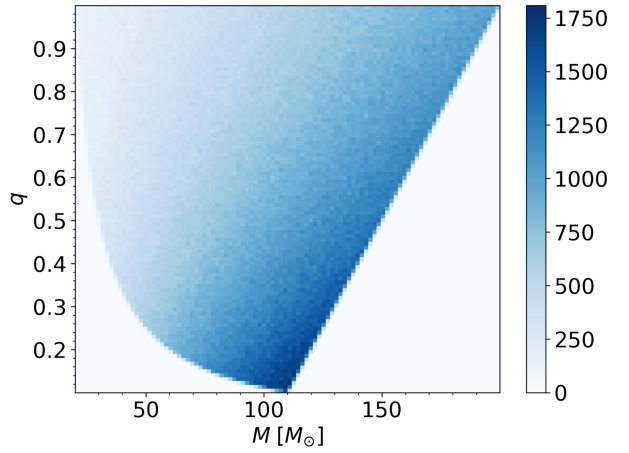
V. SIMULATIONS AND RESULTS

A. Training dataset

Since likelihood-free inference with normalizing flows relies on simulated training data samples that must reflect the properties of real ones, the simulation of the training dataset is one of the most delicate operations of this work. The dataset is made up by joint samples $\mathcal{D}=\{\boldsymbol{\theta}_{\mathrm{CE}}^{(i)},\mathbf{s}^{(i)}\}_{i=1}^{N}$ where $\boldsymbol{\theta}_{\mathrm{CE}}$ are the close encounter gravitational wave signal parameters and $\mathbf{s}$ is the corresponding strain time series sampled at 2048 Hz. This sampling frequency implies a Nyquist frequency $f_{\mathrm{nyq}}=1024$ Hz large enough to capture the frequency spectrum of CE BBH, whose peak frequency is in the band 10–100 Hz. We prepared a dataset of $N=5\times10^{6}$ samples, being the best compromise between an accurate inference and computational training cost. The first step in generating the dataset is the sampling of $\boldsymbol{\theta}_{\mathrm{CE}}$ from prior distributions. Those parameters are then fed into the effective fly-by model to produce plus and cross template polarizations $h_{+,\times}(t)$. Depending on the source sky coordinates, the template is afterward projected onto Advanced LIGO and Virgo detectors. This simulated signal is embedded into 8 seconds of Gaussian colored noise sampled from the reference O3a amplitude spectral density and saved in a hierarchical data format file. We allowed the amplitude spectral density to vary for each simulated event in order to reproduce the nonstationarity of background noise. In this work, we have not included transient noises like glitches since the capability of analyzing the time series of three detectors simultaneously automatically rejects local sources of noise. This whole procedure is parallelized, therefore significantly reducing the simulation time to $\mathcal{O}(10\ \mathrm{h})$ on a AMD EPYC 7301CPU with 32 cores / 64 threads.



(a)



(b)

FIG. 7. Prior distributions over the mass parameters. (a): uniform prior over m_i with the condition $m_2 \leq m_1$. This is the prior implemented in the simulations. (b): The same mass prior but in terms of total mass $M = m_1 + m_2$ and mass ratio $q = m_2/m_1$. Instead of the two mass components, HYPERION makes inference on M and q.

The prior distributions over $\boldsymbol{\theta}_{\rm CE}$ are listed in Table I. These population parameters can be grouped into the following categories:

Mass components: we adopted a uniform prior over $m_{1,2}$ (Fig. 7). As the strain amplitude in the EFB-T model Eq. (3) scales with the total mass M, the model makes inference on $M = m_1 + m_2$ and $q = m_2/m_1$. The condition $m_2 \leq m_1$ reduces the number of effective simulations and hence computational resources;

EFB-T parameters: namely the eccentricity e_0, semilatus rectum $\bar{p}_0$ (normalized with the total mass M) and time of peri-astron passage δt_p with respect to a reference GPS time. The ranges for these parameters are chosen for the template waveforms to provide the highest match with Numerical Relativity. In particular, we adopt the same prior choices of [20];

Gravitational wave localization parameters: the sky angles α (RA) and δ (DEC) whose prior is chosen to be uniform over the sphere. For the luminosity distance d_L we chose the range 100 Mpc–2 Gpc. We opted for a uniform prior to produce a more balanced dataset. It is worthwhile to note also that, since from Eq. (3) $h(t) \propto M^2/d_L$ a biased estimate of d_L could indeed introduce a bias also in the estimate of M;

Other gravitational wave parameters: additional parameters relevant for the simulations of the gravitational wave emission are the GPS time at which the event occurs, which is fixed for all the simulations, polarization angle ψ and inclination angle ι between the orbital angular momentum and the line of sight. For the last two, we adopt standard physical priors. At the moment, these parameters are not included in the inference process.

B. Training procedure

We trained the model for 250 training epochs, each one ending after the flow has been optimized over 1000 batches made of 512 samples.

We used the ADAM optimizer [69] with an initial learning rate of 10^{-4}. During the training, 10% of the dataset was reserved for validation, and the learning rate was reduced by 50% after 10 epochs without validation loss improvements. Before training, the training dataset is completely simulated and preprocessed. In particular, during the preprocessing phase, the strain is whitened and cropped to one second. No highpass filter was applied to avoid the risk of cutting out relevant signal frequencies. During training, we did not apply any augmentation except for the time of periastron passage δt_p, which is randomly drawn from the prior (Table I) for any training sample loaded into the GPU. The relative strain time series is rolled accordingly. Given the short duration of the signal, there is no risk for it to get too close to the time series edges.

TABLE I. Prior distributions of the simulated BBH CE population. The first set of parameters is the one over which HYPERION makes inference, while the rest enters only in the simulation phase. α and δ are right ascension and declination respectively.

$\theta_{\rm CE}$	Distribution	Minimum	Maximum
m_1 [$M_\odot$]	Uniform	10	100
$m_2 \leq m_1$ [$M_\odot$]	Uniform	10	100
$\bar{p}_0$	Uniform	13	25
d_L [Mpc]	Uniform	100	2000
e_0	Uniform	0.85	0.95
α	Uniform	0	2π
δ	cos	$-\pi/2$	$\pi/2$
δt_p [s]	uniform	−0.25	0.25
ψ	Uniform	0	π
ι	sin	0	π
GPS time	Fixed	1370692818.0	

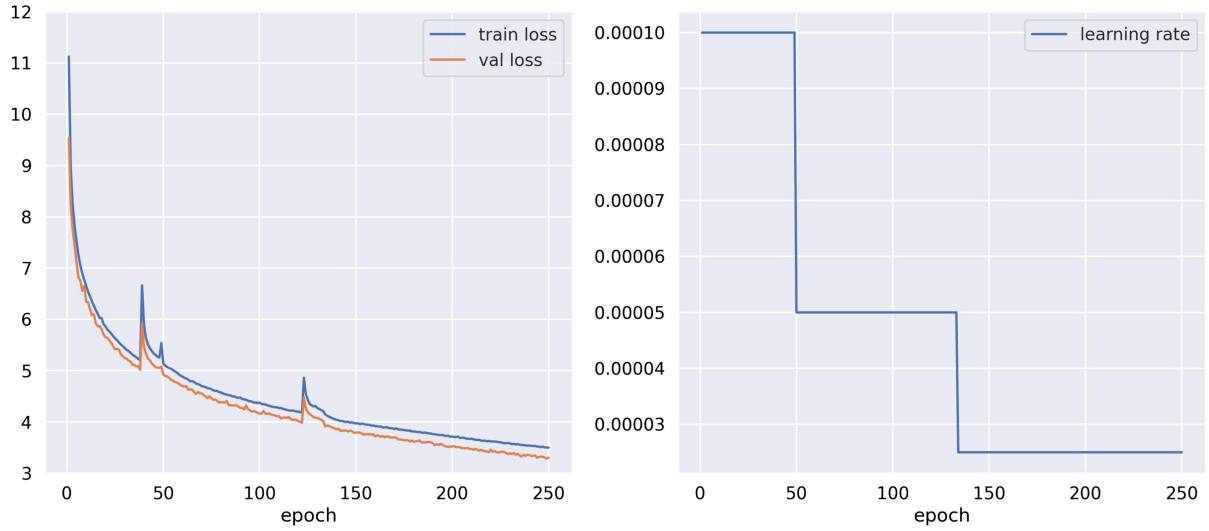


FIG. 8. Plot of the training history. Left: training and validation loss over the 250 training epochs. The close agreement between the two indicates no issue of overfitting. Right: learning rate schedule during training.

This augmentation has proven to be quite effective at reducing overfitting. The training history is shown in Fig. 8. As both the training and validation loss are in close agreement, we conclude there is no sign of overfitting. The value of 512 for the batch size is an optimal compromise between the training stability and final model accuracy. Moreover, the usage of only 1000 batches for optimization during each epoch ensures a good covering of the training sample's parameter space without the model having seen the whole dataset. This strategy is similar to the one adopted in [70].

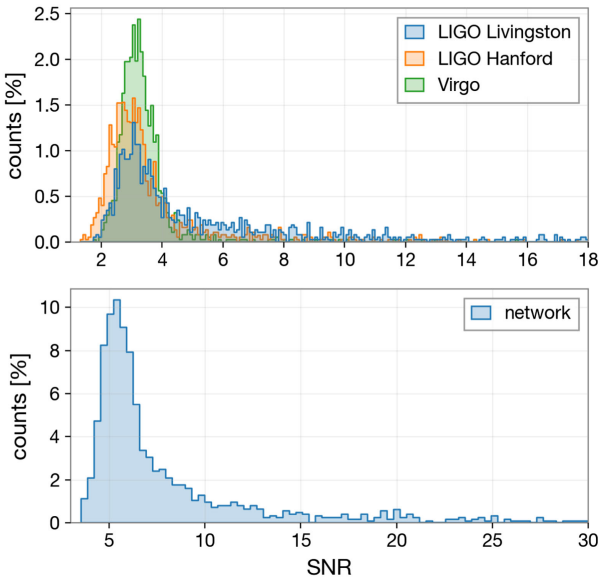


FIG. 9. SNR distribution of the signals in the Test Dataset. Top: SNR distribution for each of the simulated detectors. Bottom: network SNR distribution.

Tuning the learning rate η was a crucial aspect of the training phase. The typical starting value of $\eta_0 = 10^{-3}$ has been demonstrated to be too large and did not allow a proper optimization. We have also tested different annealing strategies, like the *cosine annealing*, although we found best results with the strategy outlined earlier. The whole training phase took around 20 hours on a Dell PowerEdge R7425 machine equipped with NVIDIA A30 GPUs.

C. Performance on parameter inference

To test the ability of HYPERION to recover $\boldsymbol{\theta}_{\mathrm{CE}}$ and its overall performance, we have simulated an additional test set. This set is composed of other 10^3 simulated signals with the same distribution as the training one. The SNR distribution of the test set is shown in Fig. 9: both for the individual detectors and for the network. The network SNR shows, in particular, a peak around a value of 5 as seen in previous works [20,60].

HYPERION's inference has been compared with the one produced by BILBY [71], adopting the DINESTY [72] sampler. We tested different hyperparameters/settings, although with minimal discrepancies in the outputs. Henceforth we will refer to the results obtained with these settings: r-walk sampling method, nlive $= 1000$, nact $= 50$, npool $= 42$.

With these parameters, 5×10^3 posterior samples were obtained in $\sim$10 hours. Using the same hardware, we produced 5×10^4 samples in 16 seconds by HYPERION running on the CPU only. When using the GPU, the same amount of samples were produced in just 0.5 seconds, improving by almost 5 orders of magnitude over standard Bayesian methods. Even on a CPU, the model can exploit, at most, hardware parallelization offered by the PYTORCH

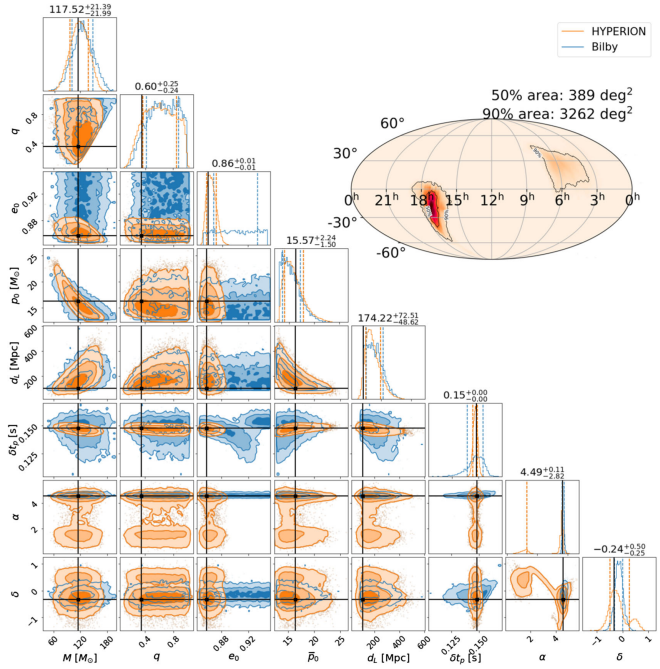


FIG. 10. Comparison of posterior samples produced by BILBY and HYPERION for a test signal with network SNR $\simeq 30$ ($d_L \simeq 100$ Mpc). The posterior for most of the parameters are well overlapping, except for eccentricity e_0 which only HYPERION is able to estimate. On the other hand, BILBY gives a slightly better estimation of the localization.

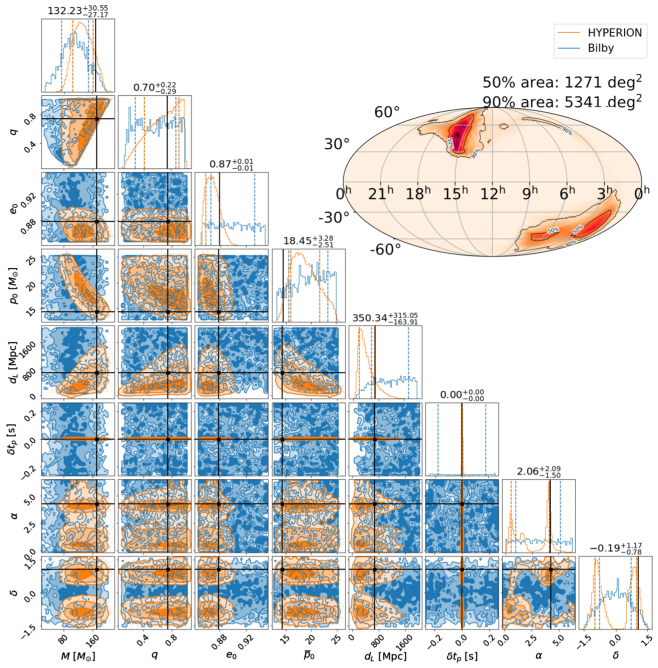


FIG. 11. Comparison of posterior samples produced by BILBY and HYPERION for a test signal with network SNR $\simeq 12$ ($d_L \simeq 700$ Mpc). In this case, only HYPERION's posteriors are informative since BILBY essentially reproduces the priors. δt_p is the better-estimated parameter. The sky-localization's posteriors show bimodality: the dominant mode is, however, the one containing the right value for (α, δ).

deep learning library. The higher inference time required by BILBY ($\mathcal{O}$(10h) is mainly due to the elevated number $\sim 10^7$–10^8 of likelihood evaluations required and the account for the autocorrelation time in the MCMC chains. In Figs. [10–12] we show corner plots comparing the obtained posteriors for some of the simulated test signals. The upper quantiles, as well as the sky-maps, refer to HYPERION. The sky-maps, in particular, are produced with a subset of 10^4 samples with the tool `ligo.skymap` [73].

VI. DISCUSSION

The results obtained by testing HYPERION on simulated data show a very promising performance when compared with traditional parameter estimation based on Bayesian inference. The agreement between the parameter's values estimated by HYPERION and Bilby (e.g., in Fig. 10) shows that NFs are a viable and robust alternative to traditional Bayesian methods since they provide the same accuracy on results but on much shorter timescales. At the same time, as shown in Fig. 12, HYPERION maintains the capability of providing informative posteriors even in the presence of low SNR signals. This can be seen, for instance, in the estimate of the total mass M in Fig. 12, where HYPERION correctly produces values peaked around the simulated values (e.g., $M \simeq 40 M_\odot$) instead of reproducing the prior distribution. Results on posteriors using Bilby suggest that low SNR signals might require additional fine-tuning of the nested sampling hyperparameters.

We note that the time shift δt_p parameter has narrow marginalized posteriors. This illustrates the efficiency of the Embedding Network and, in particular, of its Convolutional layers, which are able to recognize CE patterns even in the lowest SNR scenarios. In fact, one of the advantages of a time domain representation is that time-related patterns are directly accessible, in opposition to a frequency domain representation in which they manifest as phase shifts. Therefore, HYPERION is able to work as a standalone detection pipeline by using the Bayes factor statistic (Sec. III F). When analyzing simulated data containing only noise, the posterior for δt_p produced by HYPERION gets excessively broad, resembling the prior, thus indicating that the embedding network found no matches with known signals in the data.

As far as sky localization is concerned, we expect CE waveforms to be more difficult to localize than longer CBC waveforms, given their shorter duration and/or lower SNR. Indeed, with a shorter signal, it becomes more difficult to estimate the relative temporal shifts between the detectors because that information is concentrated in time. As a consequence, sky localization area increases with lower SNR or waveforms peaked at lower frequencies, as in Fig. 12. Although this aspect affects both standard methods

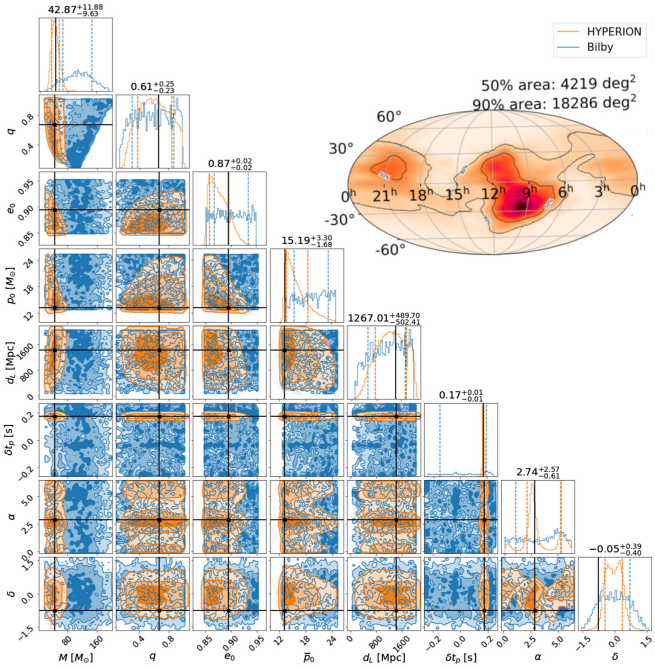


FIG. 12. Comparison of posterior samples produced by BILBY and HYPERION for a test signal with network SNR $\simeq 6$ ($d_L \simeq 1400$ Mpc). In this case, only HYPERION's posteriors are informative since BILBY reproduces the priors. The estimate of M is emblematic as BILBY completely misses the right value, which is correctly estimated by HYPERION. The greater sky-localization area can be due to this signal peaking at lower frequencies where the sensitivity is worse.

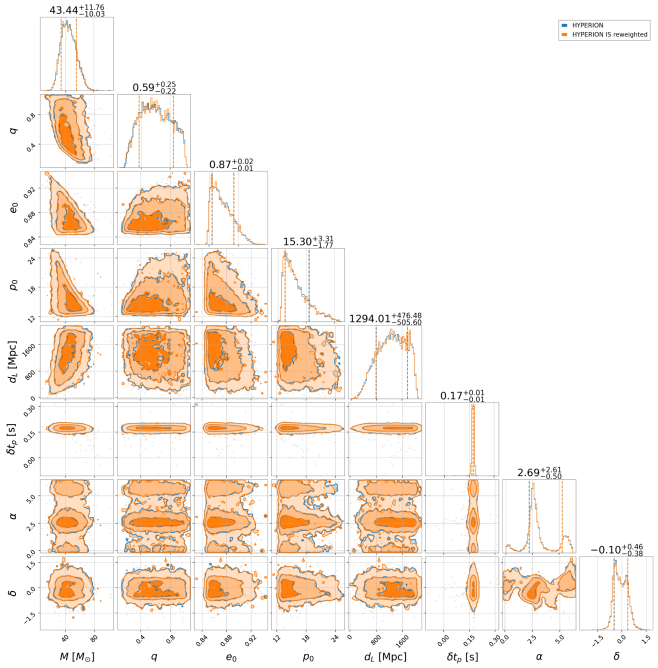


FIG. 13. Comparison between the posterior inferred by HYPERION and the importance-reweighted posterior for the test sample of Fig. 12.

and HYPERION, we notice that the latter provides better performance on localization for low SNR signals. This can be interpreted as proof of the efficiency of the localization CNN block in HYPERION's embedding network.

It can be further noticed that those posteriors (in particular the right ascension α) show multimodality, which is related to periodicity in coordinates that induces a degeneracy for values near 0 and 2π. This multimodality is also a manifestation of the ability of NFs to model complicated distributions.

To further validate HYPERION's results we reweighted the posteriors with importance sampling, with the method described in Sec. III F and compared the two distributions. A metric for the inferred posterior's goodness can be defined as $\epsilon = \frac{1}{n}(\sum_i w_i)^2/(\sum_i w_i^2) \in (0, 1]$ (*sample efficiency*) [74] where w_i are the importance weights and n the total number of posterior samples. We show an example in Fig. 13 for which $\epsilon \approx 0.8$. We obtain similar results also for other test samples. The high efficiency can be justified by the fact that the test samples comes from the same distribution as the training ones (i.e., there are no OOD samples).

Although the results on the test set suggest the good performance of HYPERION, it is crucial to provide more accurate metrics to assess the power of this approach. Given the probabilistic and Bayesian nature of the model, a suitable test for its accuracy is the probability–probability plot. This is a test used in Bayesian data analysis and widely adopted in the context of gravitational waves. The idea behind this test is to give a frequentist interpretation of Bayesian credible levels for the 1D marginalized posterior distributions. As an example, given a $CL = 0.8$, for an optimal model, it means that in the 80% of the cases, the true parameter value will lie in an interval that encloses

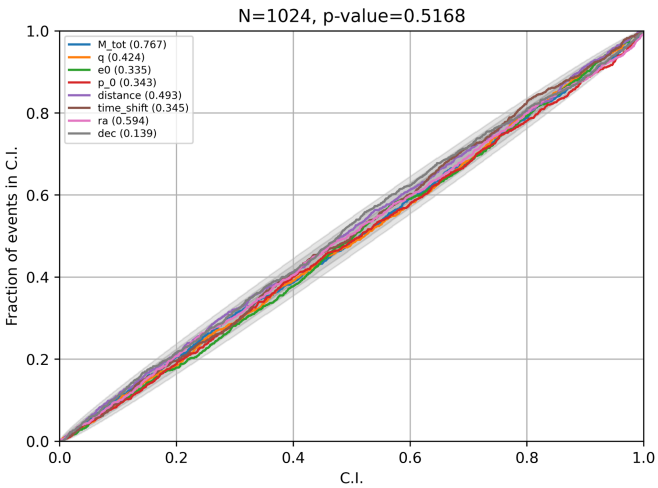


FIG. 14. Probability-probability plot for a set of 1024 posterior evaluations from the test set. Each cumulative distribution lines up pretty well along the diagonal with a spread limited within the 2σ (gray regions) for almost all the C.L. interval. In the legend is also reported for each $\boldsymbol{\theta}_{\mathrm{CE}}$ the KS statistics result. This plot has been made with a BILBY built-in function.

80% of the posterior probability, regardless of the skewness of the distribution.

To perform the test, we first drew a set of N data samples from the test set. For each of them we computed the posteriors and determined the percentile score of the true $\boldsymbol{\theta}_{\rm CE}$ parameter values in each marginalized posterior. We then took the cumulative distribution (CDF) for each of the $\boldsymbol{\theta}_{\rm CE}$. As the optimal case is represented by the percentiles being distributed according to a uniform one $\mathcal{U}(0,1)$, we tested whether the CDFs lay on the diagonal.

Figure 14 shows the result of the PP test for a set of $N = 1024$ draws. It is possible to notice that all the CDFs are well distributed along the diagonal with minimal spread limited within 2σ for almost all the C.L. intervals. To quantify how close the percentile distributions are to a uniform, a two-tail Kolmogorov-Smirnov test is performed. The output p–values are shown in Fig. 14 as well. For each of the $\boldsymbol{\theta}_{\rm CE}$, the p–values are greater than 0.1 with the combined one $\simeq 0.5$, thus implying a good recovery of the parameters. Assuming a confidence level of 95% (threshold at $\alpha = 0.05$), it is therefore not possible to reject the null hypothesis that the obtained CDFs are drawn from a uniform distribution since $p > \alpha$. The parameter with the highest p–value is the total mass M reaching $\simeq 0.7$, which does not surprise given the strong dependency of the effective fly-by waveforms on it.

One of the main differences between a normalizing flow model such as HYPERION and standard methods is that it does not use Markov chains. MCMC algorithms have to account for correlation within the chains by thinning them. This has an impact on the efficiency since the number of effective samples is reduced, or equivalently, to obtain the same $N_{\rm eff}$ samples longer chains need to be produced (see Sec. 3 of [75]). On the contrary, NFs are able to draw a set of N independent samples directly, and to prove it we computed the autocorrelation time. In particular, $\hat{\tau}_\theta$ is determined for each of the $\boldsymbol{\theta}_{\rm CE}$ set of samples with

$$\hat{\tau}_\theta = 1 + 2\sum_{\tau=1}^{M} \hat{c}_\theta(\tau) \tag{29}$$

where $\hat{c}(\tau)$ is the autocorrelation function computed with the fast Fourier transform algorithm, and M is the first τ value for which the autocorrelation exceeds a threshold value ($\hat{c}_\theta(\tau) < 0.01$).

Applying Eq. (29) the estimated autocorrelation time is $\hat{\tau}_\theta = 3 \;\forall\; \theta$: the smallest amount possible. The outlined procedure has been repeated for several different posteriors with no changes in the results. This hence indicates that all the posterior samples produced by HYPERION are valid.

We address now the major limitations of this work and how they can be alleviated in the future. Being this work a proof of concept in the analysis of close encounters, we chose limited prior bounds for the simulations. However, extended simulations can be carried out anytime. The training dataset size can be, therefore, accordingly increased, provided that it is possible to account for the higher training time. Besides, our simulations assumed Gaussian and stationary noise. By considering also artifacts like nonstationarity and/or glitches in the simulations, this inference scheme can be made even more robust.

VII. CONCLUSIONS

In this work, we introduced HYPERION, a deep learning-based pipeline to detect and perform Bayesian parameter estimation on gravitational wave signals produced by binary close encounters. No firm detection of gravitational waves from close encounters has been achieved so far, making these sources particularly interesting to broaden our view of the gravitational wave Universe. Detecting and measuring parameters of close encounters could, therefore, help to shed light on the dynamical formation channels of compact binaries and explain the observed population. Furthermore, their detection would confirm the expectations of a sub-population of compact binaries merging with non-null eccentricities.

Moreover, their low-latency detection would allow the trigger of electromagnetic follow-up observations necessary to study a potential electromagnetic counterpart as well as the surrounding environment. Detecting close encounters is difficult because of their intrinsic low signal-to-noise ratio, which makes them a hard target for current interferometers. Moreover, the short duration of the expected gravitational wave signal impacts the capability to estimate the sky coordinates and other parameters. Deep learning is a promising tool for fast analysis of gravitational wave data that could constitute a viable approach for the study of this particular source. Since the standard methods for parameter estimation are based on a Bayesian framework, we explored the application of probabilistic machine learning. In particular, we focused on Normalizing Flows, an emerging machine-learning technique that is able to infer posterior distributions on very short timescales. Compared to other methods, such as MCMC, that require many likelihood computations, NFs introduce a faster posterior sampling based on a likelihood-free approach.

The architecture of HYPERION consists of two main parts: an embedding network whose goal is to extract features from the strain time series collected by a network of ground-based interferometers and an affine coupling flow for the quick reconstruction of the posterior distribution and the estimation of the source parameters. The training of HYPERION pipeline was carried out adopting the effective fly-by waveforms on a set of $\sim 5 \times 10^6$ simulated signals, obtaining extremely promising results on the test set. The value of the reconstructed parameters is consistent with the simulated values even in low signal-to-noise ratio cases. Furthermore, the HYPERION pipeline is ~ 5 orders of magnitudes faster than traditional algorithms, providing the reconstruction of the posterior distribution on timescales of 0.5 s instead of ~ 10 h.

These results show that the NF-based approach is a viable and robust strategy for real-time detection and parameter estimation of signals from close encounters, also enabling electromagnetic follow-up campaigns.

There are several other prospects about how this work might be extended or improved in the future. In this work, we focused in particular on CE signals from binary black holes as they are the most likely to be observed with the current generation detector, both in terms of SNR and expected rates. Nevertheless, CE emission is expected also from systems containing neutron stars, and they constitute a potential source for ground-based third-generation detectors like Einstein Telescope or Cosmic Explorer [76,77], or spaceborne missions like LISA [78]. Future work will include the analysis of repeated bursts from multiple periastron encounters, allowing to track the evolution of orbital parameters during the inspiral phase.

The deep learning method presented in this work will permit rapid systematic searches for transients produced by close encounters, with the exciting possibility of detecting these signals and exploring the formation scenarios of binary compact systems in the Universe. Furthermore, this inference scheme is not limited to gravitational waves emitted by close encounters. With minimal changes, e.g., by employing a different waveform model during training and/or changing its hyperparameters, HYPERION can be adapted to search for other kinds of sources, e.g., other kinds of burst-like signals.

ACKNOWLEDGMENTS

The authors gratefully acknowledge NVIDIA Corporation for donating the two A30 GPU used in this work. L. P. acknowledges the support of the PhD scholarship in Physics "High Performance Computing and Innovative Data Analysis Methods in Science" (Cycle XXXVIII, Ministerial Decree no. 351/2022) and received funding from the European Union Next-Generation EU—National Recovery and Resilience Plan (NRRP)—MISSION 4 COMPONENT 1, INVESTMENT N.4.1—CUP N.I51J22000630007. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

[1] B. Abbott *et al.*, Observation of gravitational waves from a binary black hole merger, Phys. Rev. Lett. **116** (2016).
[2] B. Abbott *et al.*, GW170817: Observation of gravitational waves from a binary neutron star inspiral, Phys. Rev. Lett. **119** (2017).
[3] B. P. Abbott *et al.*, Multi-messenger observations of a binary neutron star merger, Astrophys. J. Lett. **848**, L12 (2017).
[4] B. P. Abbott *et al.*, Gravitational waves and gamma-rays from a binary neutron star merger: GW170817 and GRB 170817A, Astrophys. J. Lett. **848**, L13 (2017).
[5] E. Pian *et al.*, Spectroscopic identification of r-process nucleosynthesis in a double neutron-star merger, Nature (London) **551**, 67 (2017).
[6] S. J. Smartt *et al.*, A kilonova as the electromagnetic counterpart to a gravitational-wave source, Nature (London) **551**, 75 (2017).
[7] D. Radice, A. Perego, F. Zappa, and S. Bernuzzi, Gw170817: Joint constraint on the neutron star equation of state from multimessenger observations, Astrophys. J. Lett. **852**, L29 (2018).
[8] B. Margalit, Multi-messenger eos constraints using binary ns mergers, Ann. Phys. (Amsterdam) **410**, 167925 (2019).
[9] B. P. Abbott *et al.*, Tests of general relativity with GW170817, Phys. Rev. Lett. **123**, 011102 (2019).
[10] B. P. Abbott *et al.*, A gravitational-wave standard siren measurement of the Hubble constant, Nature (London) **551**, 85 (2017).
[11] R. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration), Gwtc-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run, Phys. Rev. X **13**, 041039 (2023).
[12] R. Abbott *et al.*, Population properties of compact objects from the second LIGO–Virgo gravitational-wave transient catalog, Astrophys. J. Lett. **913**, L7 (2021).
[13] The LIGO Scientific Collaboration, the Virgo Collaboration, the KAGRA Collaboration *et al.*, The population of merging compact binaries inferred using gravitational waves through GWTC-3, Phys. Rev. X **13**, 011048 (2023).
[14] M. Zevin, C. Pankow, C. L. Rodriguez, L. Sampson, E. Chase, V. Kalogera, and Frederic A. Rasio, Constraining formation models of binary black holes with gravitational-wave observations, Astrophys. J. **846**, 82 (2017).
[15] J. Samsing, Eccentric black hole mergers forming in globular clusters, Phys. Rev. D **97** (2018).
[16] C. L. Rodriguez, P. Amaro-Seoane, S. Chatterjee, K. Kremer, F. A. Rasio, J. Samsing, C. S. Ye, and M. Zevin, Post-newtonian dynamics in dense star clusters: Formation, masses, and merger rates of highly-eccentric black hole binaries, Phys. Rev. D **98** (2018).
[17] S. Bini, S. Tiwari, Y. Xu, L. Smith, M. Ebersold, G. Principe, M. Haney, P. Jetzer, and G. A. Prodi, Search for hyperbolic encounters of compact objects in the third LIGO-Virgo-KAGRA observing run, Phys. Rev. D **109**, 042009 (2024).
[18] G. Morrás, J. García-Bellido, and S. Nesseris, Search for black hole hyperbolic encounters with gravitational wave detectors, Phys. Dark Universe **35**, 100932 (2022).
[19] W. Wei, E. A. Huerta, M. Yun, N. Loutrel, M. A. Shaikh, P. Kumar, R. Haas, and V. Kindratenko, Deep learning with

quantized neural networks for gravitational-wave forecasting of eccentric compact binary coalescence, Astrophys. J. **919**, 82 (2021).
[20] N. Sorrentino, Searching for gravitational waves from binary close encounters: A deep learning analysis approach, Ph.D. thesis, Università di Pisa, 2023.
[21] S. R. Green, C. Simpson, and J. Gair, Gravitational-wave parameter estimation with autoregressive neural network flows, Phys. Rev. D **102**, 104057 (2020).
[22] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, Real-time gravitational wave science with neural posterior estimation, Phys. Rev. Lett. **127**, 241103 (2021).
[23] N. Giacobbo and M. Mapelli, The progenitors of compact-object binaries: Impact of metallicity, common envelope and natal kicks, Mon. Not. R. Astron. Soc. **480**, 2011 (2018).
[24] D. Gerosa, E. Berti, R. O'Shaughnessy, K. Belczynski, M. Kesden, D. Wysocki, and W. Gladysz, Spin orientations of merging black holes formed from the evolution of stellar binaries, Phys. Rev. D **98**, 084036 (2018).
[25] C. L. Rodriguez, M. Zevin, C. Pankow, V. Kalogera, and F. A. Rasio, Illuminating black hole binary formation channels with spins in Advanced LIGO, Astrophys. J. Lett. **832**, L2 (2016).
[26] M. Mapelli, Formation channels of single and binary stellar-mass black holes, in *Handbook of Gravitational Wave Astronomy* (Springer, Singapore, 2021), pp. 1–65.
[27] B. M. Ziosi, M. Mapelli, M. Branchesi, and G. Tormen, Dynamics of stellar black holes in young star clusters with different metallicities—II. Black hole-black hole binaries, Mon. Not. R. Astron. Soc. **441**, 3703 (2014).
[28] L. Gondán, B. Kocsis, P. Raffai, and Z. Frei, Eccentric black hole gravitational-wave capture sources in galactic nuclei: Distribution of binary parameters, Astrophys. J. **860**, 5 (2018).
[29] M. Gröbner, W. Ishibashi, S. Tiwari, M. Haney, and P. Jetzer, Binary black hole mergers in AGN accretion discs: Gravitational wave rate density estimates, Astron. Astrophys. **638**, A119 (2020).
[30] M. Zevin, J. Samsing, C. Rodriguez, C.-J. Haster, and E. Ramirez-Ruiz, Eccentric black hole mergers in dense star clusters: The role of binary–binary encounters, Astrophys. J. **871**, 91 (2019).
[31] J. Samsing, M. MacLeod, and E. Ramirez-Ruiz, The Formation of eccentric compact binary inspirals and the role of gravitational wave emission in binary-single stellar encounters, Astrophys. J. **784**, 71 (2014).
[32] J. Samsing and D. J. D'Orazio, Black hole mergers from globular clusters observable by LISA I: Eccentric sources originating from relativistic n-body dynamics, Mon. Not. R. Astron. Soc. **481**, 5445 (2018).
[33] K. Kremer *et al.*, Post-Newtonian dynamics in dense star clusters: Binary black holes in the LISA band, Phys. Rev. D **99** (2019).
[34] P. Amaro-Seoane *et al.*, Laser interferometer space antenna, arXiv:1702.00786.
[35] J. Samsing, D. J D'Orazio, K. Kremer, C. L. Rodriguez, and A. Askar, Single-single gravitational-wave captures in globular clusters: Eccentric deci-hertz sources observable by decigo and tian-qin, Phys. Rev. D **101**, 123010 (2020).
[36] D. J. D'Orazio and J. Samsing, Black hole mergers from globular clusters observable by LISA II. Resolved eccentric sources and the gravitational wave background, Mon. Not. R. Astron. Soc. **481**, 4775 (2018).
[37] N. Loutrel, Repeated bursts: Gravitational waves from highly eccentric binaries, in *Handbook of Gravitational Wave Astronomy* (Springer, Singapore, 2021), pp. 1–35.
[38] N. Loutrel, N. Yunes, and F. Pretorius, Parametrized post-Einsteinian framework for gravitational wave bursts, Phys. Rev. D **90**, 104010 (2014).
[39] J. N. Arredondo and N. Loutrel, Neutron stars in the effective fly-by framework: f-mode re-summation, Classical Quantum Gravity **38**, 165001 (2021).
[40] D. Tsang, Shattering flares during close encounters of neutron stars, Astrophys. J. **777**, 103 (2013).
[41] I. Bartos, B. Kocsis, Z. Haiman, and S. Márka, Rapid and bright stellar-mass binary black hole mergers in active Galactic nuclei, Astrophys. J. **835**, 165 (2017).
[42] B. McKernan, K. E. S. Ford, I. Bartos, M. J. Graham, W. Lyra, S. Marka, Z. Marka, N. P. Ross, D. Stern, and Y. Yang, Ram-pressure stripping of a kicked hill sphere: Prompt electromagnetic emission from the merger of stellar mass black holes in an AGN accretion disk, Astrophys. J. Lett. **884**, L50 (2019).
[43] J. García-Bellido, S. Jaraba, and S. Kuroyanagi, The stochastic gravitational wave background from close hyperbolic encounters of primordial black holes in dense clusters, Phys. Dark Universe **36**, 101009 (2022).
[44] H. Gil Choi *et al.*, Importance of eccentricities in parameter estimation of compact binary inspirals with decihertz gravitational-wave detectors, arXiv:2210.09541.
[45] W. Guo, D. Williams, I. Siong Heng, H. Gabbard, Y.-B. Bae, G. Kang, and Z.-H. Zhu, Mimicking mergers: Mistaking black hole captures as mergers, Mon. Not. R. Astron. Soc. **516**, 3847 (2022).
[46] R. Gold and B. Brügmann, Eccentric black hole mergers and Zoom-Whirl behavior from elliptic inspirals to hyperbolic encounters, Phys. Rev. D **88**, 064051 (2013).
[47] B. C. Stephens, W. E. East, and F. Pretorius, Eccentric black-hole-neutron-star mergers, Astrophys. J. **737**, L5 (2011).
[48] N. Yunes, K. G. Arun, E. Berti, and C. M. Will, Post-circular expansion of eccentric binary inspirals: Fourier-domain waveforms in the stationary phase approximation, Phys. Rev. D **80**, 084001 (2009).
[49] B. Moore and N. Yunes, A 3pn fourier domain waveform for non-spinning binaries with moderate eccentricity, Classical Quantum Gravity **36**, 185003 (2019).
[50] N. Loutrel, Analytic waveforms for eccentric gravitational wave bursts, Classical Quantum Gravity **37**, 075008 (2020).
[51] P. C. Peters and J. Mathews, Gravitational radiation from point masses in a keplerian orbit, Phys. Rev. **131**, 435 (1963).
[52] G. Papamakarios *et al.*, Normalizing flows for probabilistic modeling and inference, arXiv:1912.02762.
[53] L. Dinh *et al.*, Nice: Non-linear independent components estimation, arXiv:1410.8516.
[54] L. Weng, Flow-based deep generative models (2018), https://lilianweng.github.io/posts/2018-10-13-flow-models/.

[55] K. Cranmer, J. Brehmer, and G. Louppe, The frontier of simulation-based inference, Proc. Natl. Acad. Sci. U.S.A. **117**, 30055 (2020).
[56] S. Kullback and R. A. Leibler, On information and sufficiency, Ann. Math. Stat. **22**, 79 (1951).
[57] R. J. E. Smith, G. Ashton, A. Vajpeyi, and C. Talbot, Massively parallel Bayesian inference for transient gravitational-wave astronomy, Mon. Not. R. Astron. Soc. **498**, 4492 (2020).
[58] B. Zackay *et al.*, Relative binning and fast likelihood evaluation for gravitational wave parameter estimation, arXiv:1806.08792.
[59] J. Veitch *et al.*, Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library, Phys. Rev. D **91**, 042003 (2015).
[60] B. D. Cheeseboro and P. T. Baker, Method for detecting highly eccentric binaries with a gravitational wave burst search, Phys. Rev. D **104**, 104016 (2021).
[61] M. Maggiore *et al.*, Science case for the Einstein telescope, J. Cosmol. Astropart. Phys. 03 (2020) 050.
[62] G. Papamakarios *et al.*, Masked autoregressive flow for density estimation, arXiv:1705.07057.
[63] D. P. Kingma *et al.*, Improving variational inference with inverse autoregressive flow, arXiv:1606.04934.
[64] L. Dinh *et al.*, Density estimation using real NVP, arXiv:1605.08803.
[65] D. P. Kingma and P. Dhariwal, Glow: Generative flow with invertible 1×1 convolutions, arXiv:1807.03039.
[66] S. Kim *et al.*, FloWaveNet: A generative flow for raw audio, arXiv:1811.02155.
[67] C. Durkan *et al.*, Neural spline flows, arXiv:1906.04032.
[68] A. Paszke *et al.*, PyTorch: An imperative style, high-performance deep learning library, arXiv:1912.01703.
[69] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.
[70] H. Gabbard *et al.*, Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy, Nat. Phys. **18**, 112 (2021).
[71] G. Ashton *et al.*, Bilby: A user-friendly Bayesian inference library for gravitational-wave astronomy, Astrophys. J. Suppl. Ser. **241**, 27 (2019).
[72] J. S. Speagle, dynesty: A dynamic nested sampling package for estimating Bayesian posteriors and evidences, Mon. Not. R. Astron. Soc. **493**, 3132 (2020).
[73] L. Singer, ligo.skymap.
[74] M. Dax, S. R. Green, J. Gair, M. Pürrer, J. Wildberger, J. H. Macke, A. Buonanno, and B. Schölkopf, Neural importance sampling for rapid and reliable gravitational-wave inference, Phys. Rev. Lett. **130** (2023).
[75] A. Sokal, in *Functional Integration: Basics and Applications* (Springer US, Boston, MA, 1997), pp. 131–192.
[76] M. Punturo *et al.*, The Einstein telescope: A third-generation gravitational wave observatory, Classical Quantum Gravity **27**, 194002 (2010).
[77] E. D. Hall *et al.*, Gravitational-wave physics with cosmic explorer: Limits to low-frequency sensitivity, Phys. Rev. D **103**, 122004 (2021).
[78] P. Amaro-Seoane *et al.*, Astrophysics with the laser interferometer space antenna, Living Rev. Relativity **26**, 2 (2023).