GROUP EQUIVARIANT NEURAL POSTERIOR ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Simulation-based inference with conditional neural density estimators is a powerful approach to solving inverse problems in science. However, these methods typically treat the underlying forward model as a black box, with no way to exploit geometric properties such as equivariances. Equivariances are common in scientific models, however integrating them directly into expressive inference networks (such as normalizing flows) is not straightforward. We here describe an alternative method to incorporate equivariances under joint transformations of parameters and data. Our method—called group equivariant neural posterior estimation (GNPE)—is based on self-consistently standardizing the "pose" of the data while estimating the posterior over parameters. It is architecture-independent, and applies both to exact and approximate equivariances. As a real-world application, we use GNPE for amortized inference of astrophysical binary black hole systems from gravitational-wave observations. We show that GNPE achieves state-of-the-art accuracy while reducing inference times by three orders of magnitude.

1 INTRODUCTION

Bayesian inference provides a means of characterizing a system by comparing models against data. Given a forward model or likelihood $p(x|\theta)$ for data x described by model parameters θ, and a prior $p(\theta)$ over parameters, the Bayesian posterior distribution is proportional to their product, $p(\theta|x) \propto p(x|\theta)p(\theta)$. Sampling techniques such as Markov Chain Monte Carlo (MCMC) can be used to build up a posterior distribution provided the likelihood and prior can be evaluated.

For models with intractable or expensive likelihoods (as often arise in scientific applications) simulation-based (or likelihood-free) inference methods offer a powerful alternative (Cranmer et al., 2020). In particular, neural posterior estimation (NPE) (Papamakarios & Murray, 2016) uses expressive conditional density estimators such as normalizing flows (Rezende & Mohamed, 2015; Papamakarios et al., 2021) to build surrogates for the posterior. These are trained using model simulations $x \sim p(x|\theta)$, and allow for rapid sampling for any $x \sim p(x)$, thereby amortizing training costs across future observations. NPE and other density-estimation methods for simulation-based inference (Gutmann & Corander, 2016; Papamakarios et al., 2019; Hermans et al., 2020) have been reported to be more simulation-efficient (Lueckmann et al., 2021) than classical likelihood-free methods such as Approximate Bayesian Computation (Sisson et al., 2018).

Training an inference network for any $x \sim p(x)$ can nevertheless present challenges due to the large number of training samples and powerful networks required. The present study is motivated by applying NPE to the problem of gravitational-wave (GW) data analysis. Here the task is to infer properties of astrophysical black-hole merger events based on GW signals observed at the LIGO and Virgo observatories on Earth. Due to the complexity of signal models, it has previously not been possible to train networks to estimate posteriors to the same accuracy as conventional likelihood-based methods (Veitch et al., 2015; Ashton et al., 2019). The GW posterior, however, is equivariant[1] under an overall change in the time of arrival of the data. It is also *approximately* equivariant under a joint change in the sky position and (by triangulation) individual shifts in the arrival times in each detector (Fig. 1). If we could constrain these parameters *a priori*, we could therefore apply time shifts to align the detector data and simplify the inference task for the remaining parameters.

[1]In physics, the term "covariant" is frequently used instead of "equivariant".

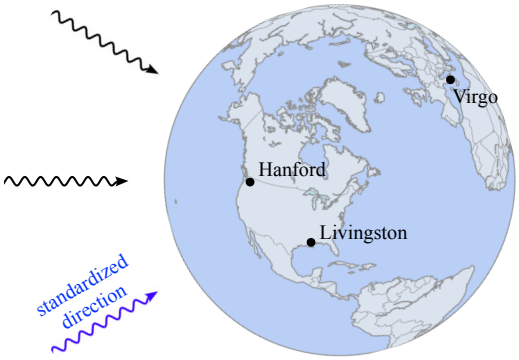

Figure 1: By standardizing the source sky position, a GW signal can be made to arrive at the same time in all three LIGO/Virgo detectors. However, since this also changes the projection of the signal onto the detectors, it defines only an *approximate* equivariance. Nevertheless, our proposed GNPE algorithm simplifies inference by simultaneously inferring *and* standardizing the incident direction.

More generally, we consider forward models with known equivariances under group transformations applied jointly to data and parameters. Our aim is to exploit this knowledge to *standardize the pose of the data*[2] and simplify analysis. The obvious roadblock is that the pose is contained in the set of parameters θ and is therefore unknown prior to inference. Here we describe *group equivariant* neural posterior estimation (GNPE), a method to self-consistently infer parameters *and* standardize the pose. The basic approach is to introduce a *proxy* for the pose—a blurred version—on which one conditions the posterior. The pose of the data is then transformed based on the proxy, placing it in a band about the standard value, and resulting in an easier inference task. Finally, the joint posterior over θ and the pose proxy can be sampled at inference time using Gibbs sampling.

The standard method to incorporating equivariances is to integrate them directly into network architectures, e.g., to use convolutional networks for translational equivariances. Although these approaches can be highly effective, they impose design constraints on network architectures. For GWs, for example, we use specialized embedding networks to extract signal waveforms from frequency-domain data, as well as expressive normalizing flows to estimate the posterior—none of which is straightforward to make explicitly equivariant. We also have complex equivariance connections between subsets of parameters and data, including approximate equivariances. The GNPE algorithm is extremely general: it is architecture-independent, it applies whether equivariances are exact or approximate, and it allows for arbitrary equivariance relations between parameters and data.

We discuss related work in section 2 and describe the GNPE algorithm in section 3. In section 4 we present a toy example with an exact translational equivariance. We show that GNPE significantly outperforms equivariance-agnostic NPE in terms of simulation efficiency, and performs on par with NPE equipped with a convolutional neural network. In section 5 we show that standard NPE does not achieve adequate accuracy for GW parameter inference, even with an essentially unlimited number of model simulations. In contrast, by exploiting physical equivariances, GNPE achieves highly accurate posteriors at a computational cost three orders of magnitude lower than bespoke MCMC approaches (Veitch et al., 2015). The present paper describes the GNPE method which we developed for GW analysis, and extends it to general equivariance transformations which make it applicable to a wide range of applications. A detailed description of its results on empirical GW data is presented in the companion physics paper (Anonymous et al., 2021). We conclude in section 6.

2 Related work

As noted above, the most common way of integrating equivariances into machine learning algorithms is to use equivariant network architectures (Krizhevsky et al., 2012; Cohen & Welling, 2016). This can be in conflict with other design considerations, such as data representation and flexibility of the architecture, and impose constraints such as locality. GNPE achieves complete separation of equivariances from these considerations, requiring only the ability to efficiently transform the pose.

Normalizing flows are particularly well suited to NPE, and there has been significant recent progress in constructing equivariant flows (Boyda et al., 2021). However, those studies have focused on joint transformation of base-space and sample-space parameters under the flow— *not* on joint transformation of data and sample-space parameters for a *conditional* flows, as we consider here.

[2]We adopt the language from computer vision by Jaderberg et al. (2015).

GNPE provides the ability for end-to-end equivariances—from data to parameters. For example, when using a convolutional embedding network in conjunction with a conditional normalizing flow, the equivariance persists up until it is passed as context to the flow, which would make it easier for the flow to learn. However, it does not fully enforce an end-to-end equivariance. This is in contrast to *invariances*, for which the above approach would be sufficient. Finally, and importantly, GNPE can also be applied if the equivariance is only *approximate*—as demonstrated in our application.

Several other approaches integrate domain knowledge of the forward model (Baydin et al., 2019; Brehmer et al., 2020) by considering a "gray-box" setting. GNPE allows us to incorporate high-level domain knowledge about approximate equivariances of forward models without requiring access to its implementation or internal states of the simulator. Rather, it can be applied to "black-box" code.

An alternative approach to incorporate geometrical knowledge into classical likelihood-free inference algorithms (e.g., Approximate Bayesian Computation, see (Sisson et al., 2018)) is by constructing (Fearnhead & Prangle, 2012) or learning (Jiang et al., 2017; Chen et al., 2021) equivariant summary statistics $s(x)$, which are used as input to the inference algorithm instead of the raw data x. However, designing equivariant summary statistics (rather than invariant ones) can be challenging, and furthermore inference will be biased if the equivariance only holds approximately.

Past studies using machine-learning techniques for amortized GW parameter inference (Gabbard et al., 2019; Chua & Vallisneri, 2020; Green & Gair, 2021; Delaunoy et al., 2020) all consider simplified problems (e.g., only a subset of parameters, a simplified posterior, or a limited treatment of detector noise). In contrast, the GNPE-based study in Anonymous et al. (2021) is the only one to treat the full amortized parameter inference problem with accuracy matching standard methods.

3 METHODS

3.1 NEURAL POSTERIOR ESTIMATION

NPE (Papamakarios & Murray, 2016) is a simulation-based inference method that directly targets the posterior. Given a dataset of prior parameter samples $\theta^{(i)} \sim p(\theta)$ and corresponding model simulations $x^{(i)} \sim p(x|\theta^{(i)})$, it trains a neural density estimator $q(\theta|x)$ to estimate $p(\theta|x)$. This is achieved by minimizing the loss

$$\mathcal{L}_{\text{NPE}} = \mathbb{E}_{p(\theta)}\mathbb{E}_{p(x|\theta)}\left[-\log q(\theta|x)\right] \tag{1}$$

across the dataset of $(\theta^{(i)}, x^{(i)})$ pairs. This maximum likelihood objective leads to recovery of $p(\theta|x)$ if $q(\theta|x)$ is sufficiently flexible. Normalizing flows (Rezende & Mohamed, 2015; Durkan et al., 2019) are a particularly expressive class of conditional density estimators commonly used for NPE.

NPE *amortizes* inference: once $q(\theta|x)$ is trained, inference is very fast for any observed data x_{o}, so training costs are shared across all observations. Furthermore, the approach treats the model as a black box, only samples from the forward model are required. Although this provides great flexibility, it does not allow one to incorporate knowledge about approximate or exact equivariances.

3.2 GROUP TRANSFORMATIONS AND POSE STANDARDIZATION

Consider for concreteness the analysis of image data, and suppose that we wish to infer the position of an object. The posterior[3] should be equivariant under translations applied jointly to the image itself and the inferred position. Namely, if we relabel the pixels of the image to apply an overall translation, then the posterior probability of the position is shifted by the same amount. Translations are composable and invertible, and there exists a trivial identity translation, so the set of translations has a natural group structure. More generally, one may be interested in other continuous transformation groups, such as rotations, scalings, etc., so in this work we will consider generic Lie groups.

[3] An equivariance of the likelihood defined by the forward model is inherited by the posterior unless explicitly broken by the prior, see App. A.1 for details. In the latter case one can still impose an invariant prior during training, and reweight posterior samples in postprocessing. We therefore consider posterior equivariances.

For a transformation group G, we denote the action of $g \in G$ on parameters θ and data x as

$$\theta \to g\theta, \tag{2}$$

$$x \to T_g x, \tag{3}$$

where T_g refers to the group representation under which the data transforms (e.g., for translations, the pixel relabeling). We are taking the convention that G is defined by its action on θ, so we do not introduce an explicit representation on parameters. The posterior is then *equivariant* under G if

$$p(\theta|x) = p(g\theta|T_g x)|\det J_g|, \qquad \forall g \in G. \tag{4}$$

Intuitively, this means that transforming the parameters with g is equivalent to transforming the observation with T_g. Here, J_g is the Jacobian of the group transformation, which arises from the change of variables rule for probability densities (see App. A.1). For translations $|\det J_g| = 1$, but this is not true in general.

We propose to incorporate equivariances by taking the G-transformed parameters to fiducial values—i.e., to standardize the pose. For our canonical example, this would mean translating the object to the center of the image. In general, θ can also include parameters unchanged under such a transformation (e.g., the color of the object). For each θ we denote the corresponding pose-standardized parameters as θ_0. These are related to θ by a group transformation g^θ (the "pose" of θ) such that $g^\theta \theta_0 = \theta$. In this terminology, standardizing the pose means to take it to the group identity element $e \in G$.

3.3 Proxy parameters for the pose

We are now presented with the basic problem that we resolve in this work. If we had access to samples from $p(g^\theta|x)$, then a neural density estimator trained on pose-standardized data would suffice for estimating the remaining parameters. However estimating the pose itself would require an additional network trained on generic (non-pose-standardized) data. The entire pipeline would therefore be limited by the latter network (as we demonstrate empirically in section 5.3).

To resolve this we make a compromise between precisely standardizing the pose and enabling parameter estimation: we transform the pose g^θ to lie within a narrow *band* around e. As we discuss in the following subsection, this gives rise to an iterative algorithm which takes successively better approximations to the parameters while transforming the pose to lie within the band.

We first augment the parameter space to include an approximation (or *proxy*) $\hat{g} \in G$ for the pose. This is defined using a kernel to blur the true pose, i.e., $\hat{g} = g^\theta \epsilon$ for $\epsilon \sim \kappa(\epsilon)$. The kernel $\kappa(\epsilon)$—a distribution over group elements $\epsilon \in G$—specifies the band to which we standardize the pose. It should be chosen to be concentrated around e, and we furthermore choose it to be symmetric. A natural choice is a simple Gaussian or uniform distribution. Then $\hat{g}$ is a close approximation to g^θ, and data x standardized with respect to $\hat{g}$ has pose within the band around e. With this definition, $p(\hat{g}|\theta) = \kappa\left((g^\theta)^{-1}\hat{g}\right)$. For translations, the group operation is addition and the identity element is the origin, so this discussion reduces to translating an image to place an object near the origin.

If $p(\theta|x)$ is equivariant under G, then it remains so when also conditioned on $\hat{g}$ (see App. A.2). The latter distribution may therefore be expressed in terms of $\hat{g}$-standardized quantities,

$$\begin{aligned} p(\theta|x, \hat{g}) &= p\left(\hat{g}^{-1}(\theta)|T_{\hat{g}^{-1}}(x), \hat{g}^{-1}\hat{g}\right)\left|\det J_{\hat{g}}^{-1}\right| \\ &\equiv p(\theta'|x')\left|\det J_{\hat{g}}^{-1}\right|. \end{aligned} \tag{5}$$

On the last line we defined θ' and x', and dropped the constant argument $\hat{g}^{-1}\hat{g} = e$. Hence, we can express the posterior conditioned on $\hat{g}$ in terms of the $\hat{g}$-standardized posterior—which is much easier to estimate. Importantly, if the equivariance is *not* exact we can still standardize the pose, but we must *explicitly* condition the $\hat{g}$-standardized posterior on $\hat{g}$ to account for equivariance-breaking effects.

3.4 Training and inference

In the general case, we train a neural density estimator $q(\theta'|x', \hat{g})$ to approximate the posterior with $\hat{g}$-standardized data. To do so, we augment the NPE maximum likelihood objective (1) with sampling over $\hat{g}$ during training,

$$\mathcal{L}_{\text{GNPE}} = \mathbb{E}_{p(\theta)}\mathbb{E}_{p(x|\theta)}\mathbb{E}_{p(\hat{g}|\theta)}\left[-\log q\left(\hat{g}^{-1}(\theta)|T_{\hat{g}^{-1}}(x), \hat{g}\right)\right]. \tag{6}$$

(Note that the logarithm causes the Jacobian in (5) to drop out upon taking the gradient, since it is independent of q.) We have explicitly conditioned q on $\hat{g}$ to allow for approximate equivariances. Exact equivariances can be strictly enforced by dropping this. Furthermore, for approximate equivariances, it may be convenient to omit the transformation $\hat{g}^{-1}$ applied to θ, which is permitted with the $\hat{g}$ conditioning.

Given a trained $q(\theta'|x', \hat{g})$ and a fixed $p(\hat{g}|\theta)$, we can self-consistently perform inference using Gibbs sampling. At inference time, samples from the joint posterior $p(\theta, \hat{g}|x)$ can be obtained by alternately sampling θ and $\hat{g}$, conditionally on the other variable and x,

$$\hat{g} \sim p(\hat{g}|x, \theta) \qquad \Longleftrightarrow \qquad \hat{g} = g^{\theta}\epsilon, \quad \epsilon \sim \kappa(\epsilon), \tag{7}$$

$$\theta \sim p(\theta|x, \hat{g}) \qquad \Longleftrightarrow \qquad \theta = \hat{g}(\theta'), \quad \theta' \sim q(\theta'|T_{\hat{g}^{-1}}(x), \hat{g}). \tag{8}$$

This procedure constructs a Markov chain with equilibrium distribution $p(\theta, \hat{g}|x)$. For convergence, the chain must be transient, aperiodic and irreducible (Roberts & Smith, 1994; Gelman et al., 2013). For sensible choices of $\kappa(\epsilon)$ the chain is both transient and aperiodic by construction. Further, irreducibility means that the entire posterior can be reached starting from any point, which should be possible even for disconnected posteriors provided the kernel is sufficiently broad. In general, burn-in truncation and thinning of the chain is required to ensure (approximately) independent samples. By marginalizing over $\hat{g}$ (i.e., ignoring it) we obtain samples from the posterior $p(\theta|x)$, as desired.[4]

Convergence of the chain also informs our choice of $\kappa(\epsilon)$. For very wide κ, $\hat{g}$ is uninformative about the pose, so sampling (8) reduces to sampling from $p(\theta|x)$, and GNPE reduces to NPE. For very narrow κ, by contrast, the pose is highly constrained but many Gibbs iterations are needed to converge. The width of κ is an additional hyperparameter, which should be chosen to accommodate this practical trade-off between convergence speed and accuracy. As a rule of thumb, the standard deviation of a typical pose posterior is a suitable starting point.

In practice, we can obtain N samples in parallel by constructing an ensemble of N Markov chains according to equations (7) and (8). We initialize these chains using samples from a second neural density estimator $q_{\text{init}}(g^{\theta}|x)$, trained using standard NPE. GNPE iterations then yield a sequence of sample sets $\{\theta_j^{(i)}\}_{i=1}^N$, $j = 0, 1, 2, \ldots$, each of which represents a distribution $Q_j(\theta|x)$ over parameters. Assuming a perfectly trained network, a GNPE iteration applied to sample set j yields an updated distribution,

$$Q_{j+1}(\theta|x) = p(\theta|x) \left[\frac{Q_j(\cdot|x) \,\bar{*}\, \kappa}{p(\cdot|x) \,\bar{*}\, \kappa} * \kappa \right] (g^{\theta}). \tag{9}$$

The "$*$" symbol denotes a group convolution, the "$\bar{*}$" symbol a combination of marginalization and group convolution. See App. A.3 for a derivation and further details. In this notation, the initialization of the pose g^{θ} in iteration 0 with q_{init} simply means setting $Q_0(\cdot|x) \,\bar{*}\, \kappa = q_{\text{init}}(\cdot|x) * \kappa$. The true posterior $p(\theta|x)$ is clearly a fixed point of this sequence, with the number of iterations to convergence determined by κ and the accuracy of the initialization network q_{init}. With κ just wide enough so that $Q_0(\cdot|x) \,\bar{*}\, \kappa \approx p(\cdot|x) \,\bar{*}\, \kappa$ for a typical observation, a single GNPE iteration suffices. Alternatively, a narrower κ can be chosen to improve accuracy of $q(\theta'|x', \hat{g})$ at the cost of additional iterations.

4 TOY EXAMPLE: DAMPED HARMONIC OSCILLATOR

As our first example, we apply GNPE to invert a simple model of a damped harmonic oscillator. The forward model gives the time-dependent position x of the oscillator, conditional on its undamped frequency ω_0, damping ratio β, and time of excitation τ. The time series x is therefore a damped sinusoid starting at τ (and zero before). Noise is introduced via a normally-distributed perturbation of the parameters $\theta = (\omega_0, \beta, \tau)$, resulting in a Gaussian posterior $p(\theta|x)$ (further details in App. B.1). The model is constructed such that the posterior is equivariant under translations in τ,

$$p(\omega_0, \beta, \tau + \Delta\tau | T_{\Delta\tau} x) = p(\omega_0, \beta, \tau | x), \tag{10}$$

so we take τ to define the pose. Note that the equivariance for this simple toy model is exact, but it could be easily turned into an approximate equivariance by, e.g., introducing a τ-dependent noise term. The prior $p(\tau)$ extends from -5 s to 0 s, so an NPE inference network would be tasked with interpreting data from oscillators excited throughout this range of times.

[4] In practice, this results only in *approximate* samples due to the asymptotic behaviour of Gibbs sampling, and a potential mismatch between the trained q and the targeted true posterior.

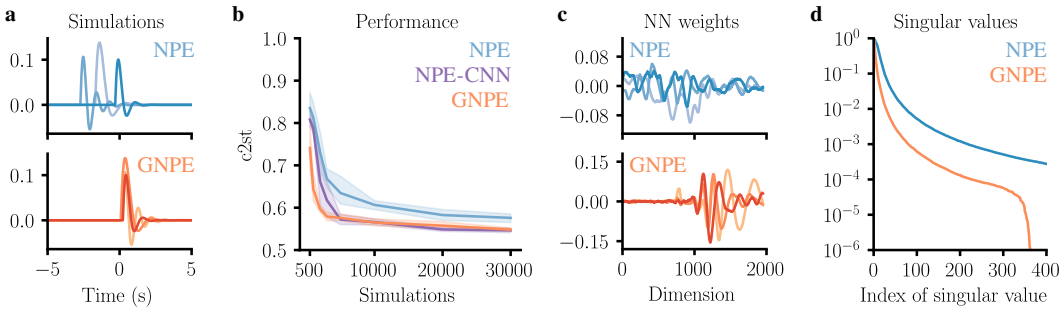

Figure 2: Comparison of standard NPE (blue) and GNPE (orange) on the damped harmonic oscillator. **a)** Three different inputs to the neural density estimators: NPE operates on the full space of simulations x, while the inputs x' to GNPE are pose-standardized. **b)** Performance in c2st score (best: 0.5, worst: 1.0). GNPE significantly outperforms equivariance-agnostic NPE, and is on par with NPE with a convolutional embedding network (purple), empirically demonstrating the importance of leveraging the translational equivariance. **c)** Example filters from the first linear layer of fully trained networks. GNPE filters more clearly capture oscillatory modes. **d)** Singular values of the training data. Inputs to GNPE x' have smaller effective dimension than raw inputs x to NPE.

For GNPE, we shift the data to align the pose within a narrow range around $\tau = 0$ (Fig. 2a). To do so we first define a proxy $\hat{\tau} = \tau + \epsilon$ for the pose, where $\epsilon \sim \kappa(\epsilon)$ is drawn from a zero-centered Gaussian kernel with standard deviation 0.1 s. We then train a neural density estimator $q(\theta'|x')$ to approximate $p(\theta'|x')$, where $\theta' \equiv (\omega_0, \beta, \tau - \hat{\tau})$ and $x' \equiv T_{-\hat{\tau}}x$ are pose-standardized. We take $q(\theta'|x')$ to be diagonal Gaussian, matching the known form of the posterior. Note that by dropping the explicit conditioning of q on $\hat{\tau}$ [cf. (6)] we enforce an exact equivariance.

For each experiment, we train until the validation loss stops decreasing. We also train a neural density estimator $q_{\text{init}}(\tau|x)$ with standard NPE on the same dataset to generate initial GNPE seeds. To generate N posterior samples we proceed as follows:

1. Sample $\tau^{(i)} \sim q_{\text{init}}(\tau|x)$, $i = 1, \ldots, N$;
2. Sample $\epsilon^{(i)} \sim \kappa(\epsilon)$, set $\hat{\tau}^{(i)} = \tau^{(i)} + \epsilon^{(i)}$, and time-translate the data, $x'^{(i)} = T_{-\hat{\tau}^{(i)}}x$;
3. Sample $\theta'^{(i)} \sim q(\theta'|x'^{(i)})$, and undo the time translation $\hat{\tau}^{(i)}$ to obtain $\theta^{(i)}$.

We repeat steps 2 and 3 until the distribution over τ converges. For this toy example and our choice of κ only one iteration is required. For further details of the implementation see App. B.2.

We evaluate GNPE on five simulations by comparing inferred samples to the ground-truth posteriors using the c2st score (Friedman, 2004; Lopez-Paz & Oquab, 2017). This score corresponds to the test accuracy of a classifier trained to discriminate samples from the target and inferred distributions, and ranges from 0.5 (best) to 1.0 (worst). As baselines we evaluate standard NPE (i) with a network architecture identical to GNPE and (ii) with a convolutional embedding network (NPE-CNN; see App. B.2). Both approaches that leverage the equivariance, GNPE (by standardizing the pose) and NPE-CNN (by using a translation-equivariant embedding network), perform similarly well and far outperform standard NPE (Fig. 2b). This underscores the importance of using equivariance-aware inference methods.

To visualize the effect of pose standardization, we show the weights of several filters from the first layer of the fully trained GNPE and NPE embedding (Fig. 2c). The NPE network is trained to interpret signals from oscillators excited at arbitrary τ, while GNPE can focus on signals starting around $\tau = 0$ (up to a small κ perturbation). This is also reflected in a reduced effective dimension of the input data to GNPE (Fig. 2d).

5 Gravitational-wave parameter inference

Gravitational waves—propagating ripples of space and time—were first detected in 2015, from the inspiral, merger, and ringdown of a pair of black holes (Abbott et al., 2016). Since that time, the two

LIGO detectors (Hanford and Livingston) (Aasi et al., 2015) as well as the Virgo detector (Acernese et al., 2015) have observed signals from over 50 coalescences of compact binaries involving either black holes or neutron stars (Abbott et al., 2019; 2021a;d). Key scientific results from these observations have included measurements of the properties of stellar-origin black holes that have provided new insights into their origin and evolution (Abbott et al., 2021b); an independent measurement of the local expansion rate of the Universe, the Hubble constant (Abbott et al., 2017); and new constraints on the properties of gravity and matter under extreme conditions (Abbott et al., 2018; 2021c).

Quasicircular binary black hole (BBH) systems are characterized by 15 parameters θ, including the component masses and spins, as well as the space-time position and orientation of the system (Tab. C.1). Given these parameters, Einstein's theory of general relativity predicts the motion and emitted gravitational radiation of the binary. The GWs propagate across billions of light-years to Earth, where they produce a time-series signal $h_I(\theta)$ in each of the LIGO/Virgo interferometers $I = \mathrm{H}, \mathrm{L}, \mathrm{V}$. The signals on Earth are very weak and embedded in detector noise n_I. In part to have a tractable likelihood, the noise is approximated as additive and stationary Gaussian. The signal and noise models give rise to a likelihood $p(x|\theta)$ for observed data $x = \{h_I(\theta) + n_I\}_{I=\mathrm{H,L,V}}$.

Once the LIGO/Virgo detection pipelines are triggered, classical stochastic samplers are typically employed to determine the parameters of the progenitor system using Bayesian inference (Veitch et al., 2015; Ashton et al., 2019). However, these methods require millions of likelihood evaluations (and hence expensive waveform simulations) for each event analyzed. Even using fast waveform models, it can take $O(\mathrm{day})$ to analyze a single BBH. Faster inference methods are therefore highly desirable to cope with growing event rates, more realistic (and expensive) waveform models, and to make rapid localization predictions for possible multimessenger counterparts. Rapid amortized methods such as NPE have the potential to transform GW data analysis. However, due to the complexity and high dimensionality[5] of GW data, it has been a challenge (Gabbard et al., 2019; Chua & Vallisneri, 2020; Green & Gair, 2021; Delaunoy et al., 2020) to obtain results of comparable accuracy and completeness to classical samplers. We now show how GNPE can be used to exploit equivariances to greatly simplify the inference problem and achieve for the first time performance indistinguishable from "ground truth" stochastic samplers—at drastically reduced inference times.

5.1 Equivariances of sky position and coalescence time

We consider the analysis of BBH systems. Included among the parameters θ are the time of coalescence t_{c} (as measured at geocenter) and the sky position (right ascension α, declination δ). Since GWs propagate at the speed of light, these are related to the times of arrival t_I of the signal in each of the interferometers.[6] Our priors (based on the precision of detection pipelines) constrain t_I to a range of ≈ 20 ms, which is much wider than typical posteriors. Standard NPE inference networks must therefore be trained on simulations with substantial time shifts.

The detector coalescence times t_I—equivalently, (t_c, α, δ)—can alternatively be interpreted as the pose of the data, and standardized using GNPE. The group G transforming the pose factorizes into a direct product of absolute and relative time shifts,

$$G = G_{\text{abs}} \times G_{\text{rel}}. \tag{11}$$

Group elements $g_{\text{abs}} \in G_{\text{abs}}$ act by uniform translation of all t_I, whereas $g_{\text{rel}} \in G_{\text{rel}}$ act by individual translation of t_{L} and t_{V}. We work with data in frequency domain, where time translations act by multiplication, $T_g x_I = e^{-2\pi i f \Delta t_I} x_I$. Absolute time shifts correspond to a shift in t_c, and are an *exact* equivariance of the posterior, $p(g_{\text{abs}}\theta | T_{g_{\text{abs}}} x) = p(\theta|x)$, where $|\det J_{g_{\text{abs}}}| = 1$. Relative time shifts correspond to a change in (α, δ) (as well as t_c). This is only an *approximate* equivariance, since a change in sky position changes the projection of the incident signal onto the detector arms, leading to a subdominant change to the signal morphology in each detector.

5.2 Application of GNPE

We use GNPE to standardize the pose within a band around $t_I = 0$. We consider two modes defined by different uniform blurring kernels. The "accurate" mode uses a narrow kernel $\kappa_{\text{narrow}} =$

[5] In our work, we analyze 8 s data segments between 20 Hz and 1024 Hz. Including also noise information, this results in 24,099 input dimensions per detector.

[6] We consider observations made in either $n_I = 2$ or 3 interferometers.

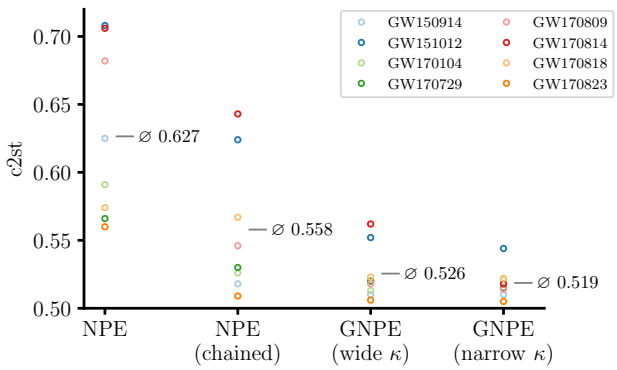

Figure 3: Comparison of estimated posteriors against LALINFERENCE MCMC for eight GW events, as quantified by c2st (best: 0.50, worst: 1.00). GNPE with a wide kernel outperforms both NPE baselines, while being only marginally slower (1 iteration $\sim$ 2 s). With a narrow kernel and 30 iterations ($\sim$ 60 s), we achieve c2st < 0.55 across all events. $\varnothing$ indicates the average across all eight events. For an alternative metric (MSE) see Fig. C.2.

$U[-1\text{ ms}, 1\text{ ms}]^{n_I}$, whereas the "fast" mode uses a wider kernel $\kappa_{\text{wide}} = U[-3\text{ ms}, 3\text{ ms}]^{n_I}$. The latter is intended to converge in just one GNPE iteration, at the cost of having to interpret a wider range of data.

We define the blurred pose proxy $\hat{g}_I \equiv t_I + \epsilon_I$, where $\epsilon_I \sim \kappa(\epsilon_I)$. We then train a conditional density estimator $q(\theta'|x', \hat{g}_{\text{rel}})$, where $\theta' = \hat{g}_{\text{abs}}^{-1}\theta$ and $x' = T_{\hat{g}^{-1}}x$. That is, we condition q on the relative time shift (since this is an approximate equivariance) and we translate parameters by the absolute time shift (since this is an exact equivariance). We always transform the data by the full time shift. We additionally train a density estimator $q_{\text{init}}(\{t_I\}_{I=\text{H,L,V}}|x)$ using standard NPE to infer initial pose estimates.

We stress that the difficulty of the inference problem (high data dimensionality, significant noise levels, complex forward model) combined with high accuracy requirements to be scientifically useful requires careful design decisions. In particular, we initialize the first layer of the embedding network with principal components of clean waveforms to provide an inductive bias to extract useful information. We further use an expressive neural-spline normalizing flow (Durkan et al., 2019) to model the complicated GW posterior structure. See App. C.2 for details of network architecture and training.

5.3 RESULTS

We evaluate performance on all eight BBH events from the first Gravitational-Wave Transient Catalog (Abbott et al., 2019) consistent with our prior bounds (component masses greater than 10 $M_\odot$). We generate reference posteriors with the standard LIGO/Virgo code LALINFERENCE (Veitch et al., 2015), which uses MCMC. We quantify the deviation between our inferred samples and the reference samples with the c2st score.

We compare performance against two baselines, standard NPE and a modified approach that partially standardizes the pose ('chained NPE'). For the latter, we use the chain rule to decompose the posterior,

$$p(\theta|x) = p(\phi, \lambda|x) = p(\phi|x, \lambda) \cdot p(\lambda|x), \tag{12}$$

where $\lambda = (t_c, \alpha, \delta)$ are the parameters required to determine the pose, and $\phi \subset \theta$ collects the remaining 12 parameters. We use standard NPE to train a flow $q(\lambda|x)$ to estimate $p(\lambda|x)$, and a flow $q(\phi|x', \lambda)$ to estimate $p(\phi|x, \lambda)$. The latter flow is conditioned on λ, which we use to standardize the pose of x. In contrast to GNPE, as noted in section 3.3, this approach is sensitive to the initial pose estimate $q(\lambda|x)$, which limits accuracy (Figs. 3 and C.7). We note that all hyperparameters of the flow and training protocol (see App. C.2) were extensively optimized on NPE, and then transferred to GNPE without modification, resulting in conservative estimates of the performance advantage of GNPE. Fast-mode GNPE converges in one iteration, whereas accurate-mode requires 30 (convergence is assessed by the JS divergence between the inferred pose posteriors from two successive iterations).

Standard NPE performs well on some GW events but lacks the required accuracy for most of them, with c2st scores up to 0.71 (Fig. 3). Chained NPE performs better across the dataset, but performs poorly on events such as GW170814, for which the initial pose estimate is inaccurate. Indeed, we find

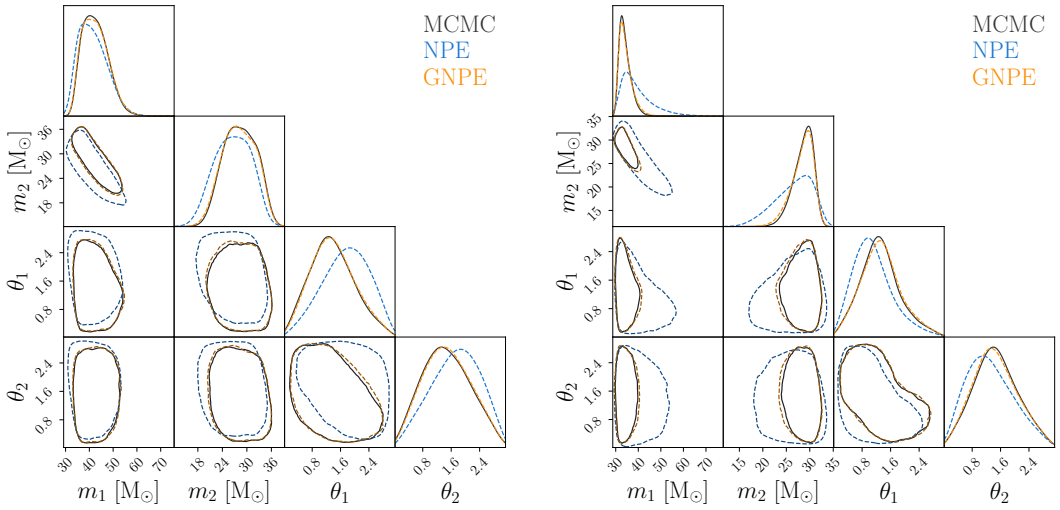

Figure 4: Corner plots for the GW events GW170809 (left) and GW170814 (right), plotting 1D marginals on the diagonal and 90% credible regions for the 2D correlations. We display the two black hole masses m_1 and m_2 and two spin parameters θ_1 and θ_2 (note that the full posterior is 15-dimensional). NPE does not accurately reproduce the MCMC posterior, while accurate-mode GNPE matches the MCMC results well. For a plot with all baselines see Fig. C.4.

that inaccuracies of that baseline can be almost entirely attributed to the initial pose estimate (Fig. C.6). Fast-mode GNPE with only one iteration is already more robust to this effect due to the blurring operation of the pose proxy (Fig. C.7). Both GNPE models significantly outperform the baselines, with accurate-mode obtaining c2st scores < 0.55 across all eight events. We emphasize that the c2st score is sensitive to any deviation between reference samples and samples from the inferred posterior. On a recent benchmark by Lueckmann et al. (2021) on examples with much lower parameter *and* data dimensions, even state-of-the-art SBI algorithms rarely reached c2st scores below 0.6. The fact that GNPE achieves scores around 0.52—i.e., posteriors which are nearly indistinguishable from the reference—on this challenging, high-dimensional, real-world example underscores the power of exploiting equivariances with GNPE.

Finally, we visualize posteriors for two events, GW170809 and GW170814, in Fig. 4. The quantitative agreement between GNPE and MCMC (Fig. 3) is visible from the overlapping marginals for all parameters displayed. NPE, by contrast, deviates significantly from MCMC in terms of shape and position. Note that we show a failure case of NPE here; for other events, such as GW170823, deviations of NPE from the reference posterior are less clearly visible.

6 Conclusions

We described GNPE, an approach to incorporate exact—and even *approximate*—equivariances under joint transformations of data and parameters into simulation-based inference. GNPE can be applied to black-box scientific forward models and any inference network architecture. The only architectural modification occurs for approximate equivariances, where additional context variables must be included to describe the pose proxy. GNPE requires similar training times compared to NPE, while the added complexity at inference time depends on the number of GNPE iterations (adjustable, but typically $O(10)$). We show with two examples that exploiting equivariances with GNPE can yield large gains in simulation efficiency and accuracy.

For the motivating problem of GW parameter estimation, GNPE achieves for the first time rapid amortized inference with results virtually indistinguishable from MCMC (Anonymous et al., 2021). This is an extremely challenging "real-world" scientific problem, with high-dimensional input data, complex signals, and significant noise levels. It combines exact and approximate equivariances, and there is no clear path to success without their inclusion along with GW-specialized architectures and expressive density estimators.

ETHICS STATEMENT

Our method is primarily targeted at scientific applications, and we do not foresee direct applications which are ethically problematic. In the context of GW analysis, we hope that GNPE contributes to reducing the required amount of compute, in particular when the rate of detections increases with more sensitive detectors in the future.

REPRODUCIBILITY STATEMENT

The experimental setup for the toy model is described in App. B.2. We also provide the code at https://tinyurl.com/wmbjajv8.

The setup for GW parameter inference is described in App. C.2. We will release a python package including scripts for the experiments carried out in section 5 upon acceptance, and will also be happy to share them with the reviewers anonymously during the rebuttal period as suggested on the ICLR FAQ.

REFERENCES

J. Aasi et al. Advanced LIGO. *Class. Quant. Grav.*, 32:074001, 2015. doi: 10.1088/0264-9381/32/7/074001.

B. P. Abbott et al. Observation of Gravitational Waves from a Binary Black Hole Merger. *Phys. Rev. Lett.*, 116(6):061102, 2016. doi: 10.1103/PhysRevLett.116.061102.

B. P. Abbott et al. A gravitational-wave standard siren measurement of the Hubble constant. *Nature*, 551(7678):85–88, 2017. doi: 10.1038/nature24471.

B. P. Abbott et al. GW170817: Measurements of neutron star radii and equation of state. *Phys. Rev. Lett.*, 121(16):161101, 2018. doi: 10.1103/PhysRevLett.121.161101.

B. P. Abbott et al. GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs. *Phys. Rev. X*, 9(3): 031040, 2019. doi: 10.1103/PhysRevX.9.031040.

R. Abbott et al. GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run. *Phys. Rev. X*, 11:021053, 2021a. doi: 10.1103/PhysRevX.11.021053.

R. Abbott et al. Population Properties of Compact Objects from the Second LIGO-Virgo Gravitational-Wave Transient Catalog. *Astrophys. J. Lett.*, 913(1):L7, 2021b. doi: 10.3847/2041-8213/abe949.

R. Abbott et al. Tests of general relativity with binary black holes from the second LIGO-Virgo gravitational-wave transient catalog. *Phys. Rev. D*, 103(12):122002, 2021c. doi: 10.1103/PhysRevD.103.122002.

R. Abbott et al. GWTC-2.1: Deep Extended Catalog of Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run. 8 2021d.

F. Acernese et al. Advanced Virgo: a second-generation interferometric gravitational wave detector. *Class. Quant. Grav.*, 32(2):024001, 2015. doi: 10.1088/0264-9381/32/2/024001.

A. Anonymous et al. Anonymous title. *arXiv preprint arXiv:2106.XXXXX*, 2021.

Gregory Ashton et al. BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy. *Astrophys. J. Suppl.*, 241(2):27, 2019. doi: 10.3847/1538-4365/ab06fc.

Atilim Güneş Baydin, Lei Shao, Wahid Bhimji, Lukas Heinrich, Lawrence Meadows, Jialin Liu, Andreas Munk, Saeid Naderiparizi, Bradley Gram-Hansen, Gilles Louppe, et al. Etalumis: Bringing probabilistic programming to scientific simulators at scale. In *Proceedings of the international conference for high performance computing, networking, storage and analysis*, pp. 1–24, 2019.

Alejandro Bohé, Mark Hannam, Sascha Husa, Frank Ohme, Michael Pürrer, and Patricia Schmidt. PhenomPv2 – technical notes for the LAL implementation. *LIGO Technical Document, LIGO-T1500602-v4*, 2016. URL https://dcc.ligo.org/LIGO-T1500602/public.

Denis Boyda, Gurtej Kanwar, Sébastien Racanière, Danilo Jimenez Rezende, Michael S Albergo, Kyle Cranmer, Daniel C Hackett, and Phiala E Shanahan. Sampling using su (n) gauge equivariant flows. *Physical Review D*, 103(7):074504, 2021.

Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer. Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences*, 117(10): 5242–5249, 2020.

Yanzhi Chen, Dinghuai Zhang, Michael U Gutmann, Aaron Courville, and Zhanxing Zhu. Neural approximate sufficient statistics for implicit models. In *Ninth International Conference on Learning Representations 2021*, 2021.

Alvin J. K. Chua and Michele Vallisneri. Learning Bayesian posteriors with neural networks for gravitational-wave inference. *Phys. Rev. Lett.*, 124(4):041102, 2020. doi: 10.1103/PhysRevLett.124.041102.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Arnaud Delaunoy, Antoine Wehenkel, Tanja Hinderer, Samaya Nissanke, Christoph Weniger, Andrew R Williamson, and Gilles Louppe. Lightning-fast gravitational wave parameter inference through neural amortization. *Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020)*, 2020.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in Neural Information Processing Systems*, 32:7511–7522, 2019.

Benjamin Farr, Evan Ochsner, Will M. Farr, and Richard O'Shaughnessy. A more effective coordinate system for parameter estimation of precessing compact binaries from gravitational waves. *Phys. Rev. D*, 90(2):024018, 2014. doi: 10.1103/PhysRevD.90.024018.

Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, 2012.

Jerome Friedman. On multivariate goodness-of-fit and two-sample testing. In *Conference on Statistical Problems in Particle Physics, Astrophysics and Cosmology*, 2004.

Hunter Gabbard, Chris Messenger, Ik Siong Heng, Francesco Tonolini, and Roderick Murray-Smith. Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy, 2019.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

Stephen R. Green and Jonathan Gair. Complete parameter inference for GW150914 using deep learning. *Mach. Learn. Sci. Tech.*, 2(3):03LT01, 2021. doi: 10.1088/2632-2153/abfaed.

Michael U Gutmann and Jukka Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *The Journal of Machine Learning Research*, 17(1):4256–4302, 2016.

Mark Hannam, Patricia Schmidt, Alejandro Bohé, Leïla Haegel, Sascha Husa, Frank Ohme, Geraint Pratten, and Michael Pürrer. Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms. *Phys. Rev. Lett.*, 113(15):151101, 2014. doi: 10.1103/PhysRevLett.113.151101.

Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with approximate likelihood ratios. In *Proceedings of the 37th International Conference on Machine Learning*, volume 98 of *Proceedings of Machine Learning Research*. PMLR, 2020.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.

Bai Jiang, Tung-yu Wu, Charles Zheng, and Wing H Wong. Learning summary statistic for approximate bayesian computation via deep neural network. *Statistica Sinica*, pp. 1595–1618, 2017.

Sebastian Khan, Sascha Husa, Mark Hannam, Frank Ohme, Michael Pürrer, Xisco Jimńez Forteza, and Alejandro Bohé. Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era. *Phys. Rev.*, D93(4):044007, 2016. doi: 10.1103/PhysRevD.93.044007.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *5th International Conference on Learning Representations, ICLR*, 2017.

Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 343–351. PMLR, 2021.

George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. In *Advances in neural information processing systems*, pp. 1028–1036, 2016.

George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pp. 837–848. PMLR, 2019.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. URL `http://jmlr.org/papers/v22/19-1028.html`.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Gareth O Roberts and Adrian FM Smith. Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. *Stochastic processes and their applications*, 49(2):207–216, 1994.

Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian computation*. CRC Press, 2018.

Alvaro Tejero-Cantero, Jan Boelts, Michael Deistler, Jan-Matthis Lueckmann, Conor Durkan, Pedro J. Gonçalves, David S. Greenberg, and Jakob H. Macke. sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505, 2020. doi: 10.21105/joss.02505.

J. Veitch et al. Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library. *Phys. Rev.*, D91(4):042003, 2015. doi: 10.1103/PhysRevD.91.042003.

A DERIVATIONS

A.1 EQUIVARIANCE RELATIONS

Consider a system with an exact equivariance under a joint transformation of parameters θ and observations x,

$$\theta \to \theta' = g\theta, \tag{13}$$

$$x \to x' = T_g x. \tag{14}$$

An invariant prior fulfills the relation

$$p(\theta) = p(\theta') \left|\det J_g\right|, \tag{15}$$

where the Jacobian J_g arises from the change of variables rule for probability distributions. It is required for the normalization to remain intact,

$$1 = \int p(\theta)\, d\theta = \int p(\theta') \left|\det J_g\right| \, d\theta = \int p(\theta')\, d\theta'. \tag{16}$$

Note that we neglect boundary effects here. A similar relation holds for an equivariant likelihood,

$$p(x|\theta) = p(x'|\theta') \left|\det J_T\right|. \tag{17}$$

Again, the Jacobian is required for the normalization,

$$1 = \int p(x|\theta)p(\theta)\, dx d\theta = \int p(x'|\theta') \left|\det J_T\right| p(\theta') \left|\det J_g\right| \, dx d\theta. \tag{18}$$

An invariant prior and an equivariant likelihood further imply for the evidence $p(x)$

$$p(x) = \int p(x|\theta)p(\theta) d\theta = \int p(x'|\theta') \left|\det J_T\right| p(\theta') \left|\det J_g\right| \, d\theta = p(x') \left|\det J_T\right|. \tag{19}$$

Combining an invariant prior with an equivariant likelihood thus leads to the equivariance relation

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x'|\theta') \left|\det J_T\right| p(\theta') \left|\det J_g\right|}{p(x') \left|\det J_T\right|} \tag{20}$$

$$= p(\theta'|x') \left|\det J_g\right| \tag{21}$$

for the posterior, where we used Bayes' theorem and equations (15), (17), and (19).

A.2 EQUIVARIANCE OF $p(\theta|x, \hat{g})$

Here we derive that an equivariant posterior $p(\theta|x)$ remains equivariant if the distribution is also conditioned on the proxy $\hat{g}$, as used in equation (5). With the definition $p(\hat{g}|\theta) = \kappa\left((g^\theta)^{-1}\hat{g}\right)$ from section 3.3, the $p(\hat{g}|\theta)$ is equivariant under joint application of $h \in G$ to $\hat{g}$ and θ,

$$p(\hat{g}|\theta) = \kappa\left((g^\theta)^{-1}\hat{g}\right) = \kappa\left((g^\theta)^{-1}h^{-1}h\hat{g}\right) = \kappa\left((hg^\theta)^{-1}h\hat{g}\right) = p(h\hat{g}|h\theta). \tag{22}$$

Here, we used that by definition a group element h directly alters the pose, $g^{h\theta} = hg^\theta$. This implies, that $p(\hat{g}|x)$ is equivariant under joint application of h and T_h,

$$\begin{aligned} p(\hat{g}|x) &= \int p(\hat{g}|\theta, x)p(\theta|x)\, d\theta \overset{(22),(4)}{=} \int p(h\hat{g}|h\theta)p(h\theta|T_h x) \left|\det J_h\right| \, d\theta \\ &= p(h\hat{g}|T_h x). \end{aligned} \tag{23}$$

in the second step we used $p(\hat{g}|\theta, x) = p(\hat{g}|\theta)$. From these relations, the equivariance relation used in equation (5) follows,

$$\begin{aligned} p(\theta|x, \hat{g}) &= \frac{p(\theta, \hat{g}|x)}{p(\hat{g}|x)} = \frac{p(\hat{g}|\theta, x)p(\theta|x)}{p(\hat{g}|x)} \overset{(22),(4),(23)}{=} \frac{p(h\hat{g}|h\theta)p(h\theta|T_h x)}{p(h\hat{g}|T_h x)} \left|\det J_h\right| \\ &= p(h\theta|T_h x, h\hat{g}) \left|\det J_h\right|. \end{aligned} \tag{24}$$

A.3 ITERATIVE INFERENCE AND CONVERGENCE

GNPE leverages a neural density estimator of the form $q(\theta|x', \hat{g})$ to obtain samples from the joint distribution $p(\theta, \hat{g}|x)$. This is done by iterative sampling as described in section 3. Here we derive equation (9), which states how a distribution $Q_j(\theta|x)$ is updated by a single GNPE iteration.

Given a distribution $Q_j^\theta(\theta|x)$ of θ samples in iteration j, we infer samples for the pose proxy $\hat{g}$ for the next iteration by (i) extracting the pose g^θ from θ (this essentially involves marginalizing over all non pose related parameters) and (ii) blurring the pose g^θ with the kernel κ, corresponding to a group convolution.[A.1] We denote this combination of marginalization and group convolution with the "$\bar{*}$" symbol,

$$Q_{j+1}^{\hat{g}}(\hat{g}|x) = \int d\theta \, Q_j^\theta(\theta|x) \kappa((g^\theta)^{-1}\hat{g}) = \left(Q_j^\theta(\cdot|x) \,\bar{*}\, \kappa\right)(\hat{g}). \tag{25}$$

For a given proxy sample $\hat{g}$, a (perfectly trained) neural density estimator infers θ with

$$p(\theta|x, \hat{g}) = \frac{p(\theta, \hat{g}|x)}{p(\hat{g}|x)} = \frac{p(\hat{g}|x, \theta)p(\theta|x)}{p(\hat{g}|x)} = p(\theta|x) \frac{\kappa((g^\theta)^{-1}\hat{g})}{(p^\theta(\cdot|x) \,\bar{*}\, \kappa)(\hat{g})}, \tag{26}$$

where we used $p(\hat{g}|\theta) = \kappa((g^\theta)^{-1}\hat{g})$. Combining (25) and (26), the updated distribution over θ samples reads

$$\begin{aligned} Q_{j+1}^\theta(\theta|x) &= \int d\hat{g} \, p(\theta|x, \hat{g}) \, Q_{j+1}^{\hat{g}}(\hat{g}|x) \\ &= \int d\hat{g} \, \left(Q_j^\theta(\cdot|x) \,\bar{*}\, \kappa\right)(\hat{g}) \, p(\theta|x) \frac{\kappa((g^\theta)^{-1}\hat{g})}{((p^\theta(\cdot|x) \,\bar{*}\, \kappa))(\hat{g})} \\ &= p(\theta|x) \int d\hat{g} \, \frac{\left(Q_j^\theta(\cdot|x) \,\bar{*}\, \kappa\right)(\hat{g})}{((p^\theta(\cdot|x) \,\bar{*}\, \kappa))(\hat{g})} \, \kappa((g^\theta)^{-1}\hat{g}) \\ &= p(\theta|x) \left(\frac{Q_j^\theta(\cdot|x) \,\bar{*}\, \kappa}{p^\theta(\cdot|x) \,\bar{*}\, \kappa} \; * \; \kappa^{(-)} \right)(\hat{g}). \end{aligned} \tag{27}$$

Here, $\kappa^{(-)}$ denotes the reflected kernel, $\kappa^{(-)}(g) = \kappa(g^{-1}) \, \forall g$. Since we choose a symmetric kernel in practice, we use $\kappa = \kappa^{(-)}$ in (9).

B TOY EXAMPLE

B.1 FORWARD MODEL

The toy model in section 4 describes the motion of a damped harmonic oscillator that is initially at rest and excited at time τ with an infinitely short pulse. The time evolution of that system is governed by the differential equation

$$\frac{d^2}{dt^2}x(t) + 2\beta\omega_0 \frac{d}{dt}x(t) + \omega_0^2 x(t) = \delta(t - \tau), \tag{28}$$

where ω_0 denotes the undamped angular frequency and β the damping ratio. The solution for the time series $x(t)$ is given by the Green's function for the corresponding differential operator and reads

$$x(t) = \begin{cases} 0, & t \leq \tau \\ e^{-\beta\omega_0(t-\tau)} \cdot \frac{\sin\left(\sqrt{1-\beta^2}\omega_0(t-\tau)\right)}{\sqrt{1-\beta^2}\omega_0}, & t > \tau. \end{cases} \tag{29}$$

This equation describes a deterministic, injective mapping between parameters $\theta = (\omega_0, \beta, \tau)$ and a time series observation x,

$$x = f(\theta). \tag{30}$$

[A.1] We define a group convolution as $(A * B)(\hat{g}) = \int dg \, A(g) B(g^{-1}\hat{g})$, which is the natural extension of a standard convolution.

This implies a likelihood $p(x|\theta) = \delta(x - f(\theta))$, and thus a point-like posterior. To showcase (G)NPE on this toy problem we introduce stochasticity by setting

$$x = f(\theta + \delta\theta) \tag{31}$$

instead. We sample $\delta\theta$ from an uncorrelated Gaussian distribution

$$\delta\theta \sim \mathcal{N}(0, \Sigma), \qquad \Sigma = \begin{pmatrix} \sigma_{\omega_0}^2 & 0 & 0 \\ 0 & \sigma_\beta^2 & 0 \\ 0 & 0 & \sigma_\tau^2 \end{pmatrix}, \tag{32}$$

with $\sigma_{\omega_0} = 0.3$ Hz, $\sigma_\beta = 0.03$ and $\sigma_\tau = 0.3$ s. Due to the injectivity of f, the posterior $p(\theta|x)$ reduces to the probability $p(\delta\theta = f^{-1}(x) - \theta)$. With a uniform prior, and neglecting boundary effects, this implies an uncorrelated Gaussian posterior $p(\theta|x)$ centered around $f^{-1}(x)$ with standard deviations as specified above. We choose this approach over, e.g., adding noise straight to observations to keep the problem as simple as possible, such that the focus remains on the comparison of GNPE and NPE. In particular, knowing that the ground truth posteriors are Gaussian, we can use a simple Gaussian density estimator.

We choose uniform priors

$$p(\omega_0) = \mathcal{U}[3, 10] \text{ Hz}, \qquad p(\beta) = \mathcal{U}[0.2, 0.5], \qquad p(\tau) = \mathcal{U}[-5, 0] \text{ s}. \tag{33}$$

The observational data x is the discretized time series in the interval $[-5, +5]$ s with 2000 evenly sampled bins. When applying time shifts with GNPE, we impose cyclic boundary conditions.

B.2 Implementation

We use a Gaussian density estimator for all methods (since we know that the true posterior is Gaussian). For NPE, we use a feedforward neural network with [128, 32, 16] hidden units and with ReLU activation functions as an embedding network. For NPE-CNN, we use a three-layer convolutional embedding network with kernel sizes [5,5,5], stride 1, [6,12,12] channels, average pooling with kernel size 7 and stride 7, and ReLU activation functions. For GNPE, we use the same architecture as for NPE for both, $q(\theta'|x')$ and $q_{\text{init}}(\tau|x)$. For further hyperparameters, we use the defaults of the sbi package (Tejero-Cantero et al., 2020).

B.3 Results

For all methods, we compute the average classifier two-sample test score (c2st) based on 10,000 samples from the estimated and the ground truth posterior for five different simulations. We then average the accuracy across 10 different seeds.

C Gravitational wave parameter inference

C.1 Forward model and amortization

The forward model mapping binary black hole parameters θ (Tab. C.1) to simulated measurements x in the detectors consists of two stages. Firstly, the waveform polarizations $h(\theta)$ for given parameters θ are computed with the waveform model IMRPhenomPv2 (Hannam et al., 2014; Khan et al., 2016; Bohé et al., 2016). Secondly, the signals are projected onto the detectors, and noise is added to obtain a realistic signal x. To a good approximation, we assume the noise to be Gaussian and stationary over the duration of a single GW signal. However, the noise spectrum, determined by the power spectral density (PSD) S_n, drifts over the duration of an observing run. To fully amortize the computational cost, we use a variety of different PSDs S_n in training, and additionally condition the inference network on S_n. At inference time, this enables instant tuning of the inference network to the PSD estimated at the time of the event, see Anonymous et al. (2021) for details. Since this conditioning on S_n has no effect on the GNPE algorithm outlined in this work, we keep it implicit in all equations.

C.2 Network architecture and training

The inference network consists of an embedding network, that reduces the high dimensional input data to a 128 dimensional feature vector, and the normalizing flow, that takes this feature vector

Table C.1: Priors for the astrophysical binary black hole parameters used to train the inference network. Priors are uniform over the specified range unless indicated otherwise. We train networks with different distance ranges for the two observing runs O1 and O2 due to the different detector sensitivities. At inference time, a cosmological distance prior is imposed by reweighting samples according to their distance.

Description	Parameter	Prior
component masses	m_1, m_2	$[10, 80]$ $\mathrm{M}_\odot$, $m_1 \geq m_2$
spin magnitudes	a_1, a_2	$[0, 0.88]$
spin angles	θ_1, θ_2, ϕ_{12}, ϕ_{JL}	standard as in Farr et al. (2014)
time of coalescence	t_c	$[-0.1, 0.1]$ s
luminosity distance	d_L	O1, 2 detectors: $[100, 2000]$ Mpc
		O2, 2 detectors: $[100, 2000]$ Mpc and $[100, 6000]$ Mpc
		O2, 3 detectors: $[100, 1000]$ Mpc
reference phase	ϕ_c	$[0, 2\pi]$
inclination	θ_{JN}	$[0, \pi]$ uniform in sine
polarization	ψ	$[0, \pi]$
sky position	α, β	uniform over sky

as input. For each detector, the input to the embedding network consists of the complex-valued frequency domain strain in the range $[20 \text{ Hz}, 1024 \text{ Hz}]$ with a resolution of 0.125 Hz, and PSD information $(10^{46} \cdot S_\text{n})^{-1/2}$ with the same binning. This results to a total of $(3 \cdot 8{,}033) = 24{,}099$ real input bins per detector. The first module of the embedding network consists of a linear layer per detector, that maps this 24,096 dimensional input to 400 components. We initialize this compression layer with PCA components of raw waveforms. This provides a strong inductive bias to the network to filter out GW signals from extremely noisy data. Note that this important step is only possible since GNPE is architecture independent—it is for instance not compatible with a convolutional neural network. Following this compression layer, we use a series of 24 fully-connected residual blocks with two layers each to compress the output to the desired 128 dimensional feature vector. We use batch normalization and ELU activation functions. Importantly, the conditioning of the flow on the proxy $\hat{g}_\text{rel.}$ is done *after* the embedding network, by concatenating $\hat{g}_\text{rel.}$ to the embedded feature vector.

Following this, we use a neural spline flow (Durkan et al., 2019) with rational-quadratic spline coupling transforms as density estimator. We use 30 such transforms, each of which is associated with 5 two-layer residual blocks with hidden dimension 512. In total, the inference network has 348 hidden layers and $1.31 \cdot 10^8$ (for two detectors) or $1.42 \cdot 10^8$ (for three detectors) learnable parameters.

We train the inference network with a data set of $5 \cdot 10^6$ waveforms with parameters θ sampled from the priors specified in table C.1, and reserve 2% of the data for validation. We pretrain the network with learning rate of $3 \cdot 10^{-4}$ for 300 epochs with fixed PSD, and finetune for another 150 epochs with learning rate of $3 \cdot 10^{-5}$ with varying PSDs. With batch size 4,096, training takes 16-18 days on a NVIDIA Tesla V100 GPU.

C.3 Results

The c2st scores between inferred posterior and the MCMC reference shown in Fig. 3 are computed using the code and default hyperparameters of Lueckmann et al. (2021). For each event, we compute the c2st score of 10,000 samples for inferred and target posterior. Fig. 3 displays the mean of the score across 5 different sample realizations, Fig. C.1 additionally shows the corresponding standard deviation. For technical reasons we use only 12 of the 15 inferred parameters; specifically we omit the geocentric time of coalescence t_c (since the reference posteriors generated with LALInference do not contain that variable) and the sky position parameters α and δ (since the NPE baseline with chain rule decomposition infers these in another basis).

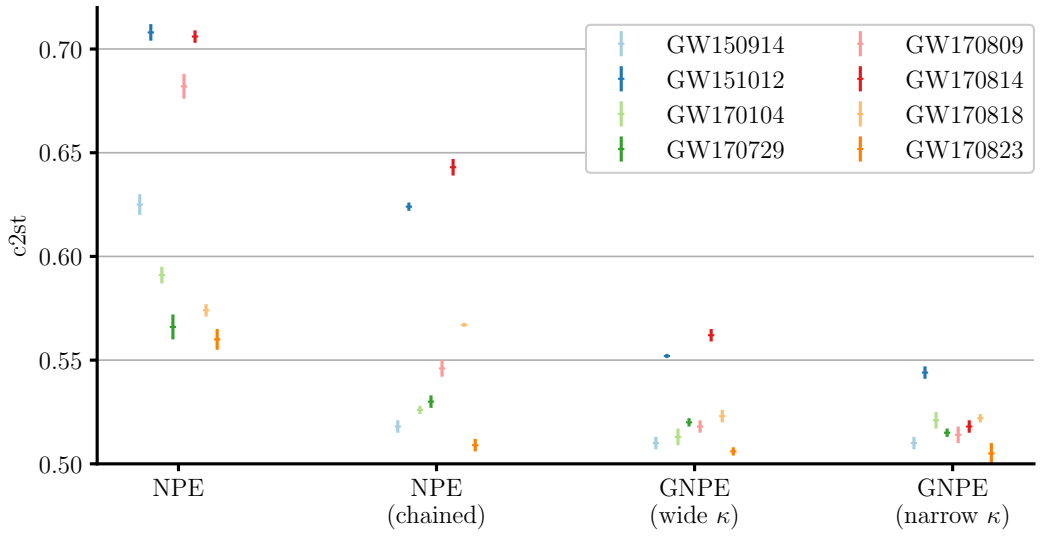

Figure C.1: c2st scores quantifying the deviation between the inferred posteriors and the MCMC reference. This is an extended version of Fig. 3.

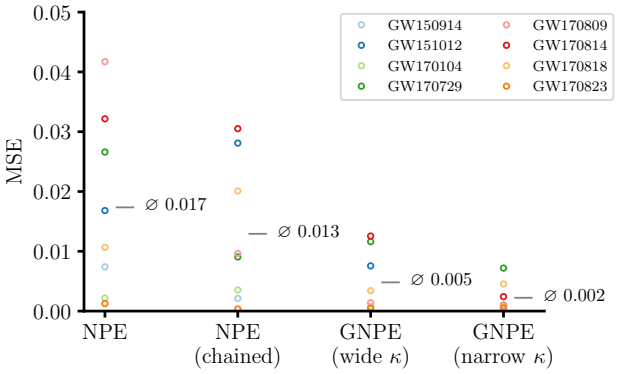

Figure C.2: Comparison of estimated posteriors against LALINFERENCE MCMC for eight GW events, as quantified by the mean squared error (MSE) of the sample means. Before computing the means, we normalize each dimension such that the prior has a standard deviation of 1. ∅ indicates the average across all eight events. GNPE with a narrow kernel consistently outperforms the baselines, which is in accordance with Fig. 3.

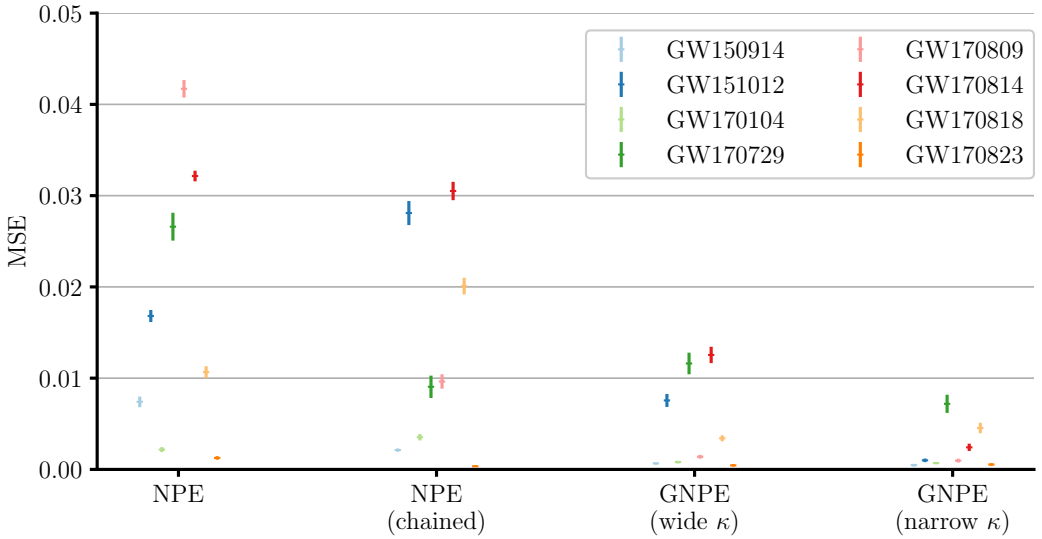

Figure C.3: MSE between inferred posteriors and MCMC reference. Extended version of Fig. C.2.

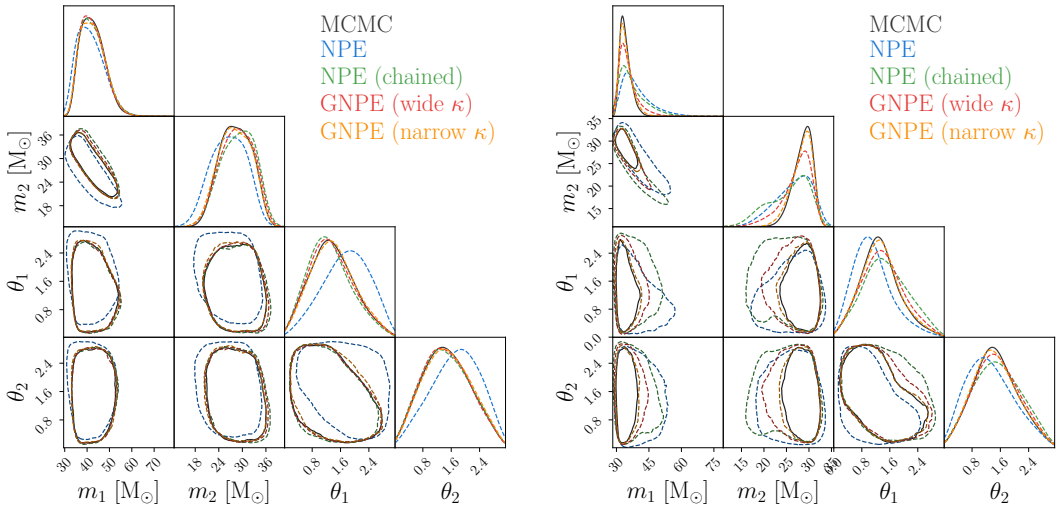

Figure C.4: Corner plots for the GW events GW170809 (left) and GW170814 (right), plotting 1D marginals on the diagonal and 90% credible regions for the 2D correlations. We display the two black hole masses m_1 and m_2 and two spin parameters θ_1 and θ_2 (note that the full posterior is 15-dimensional). This extends Fig. 4 by also displaying the results from chained NPE and GNPE with wide κ.

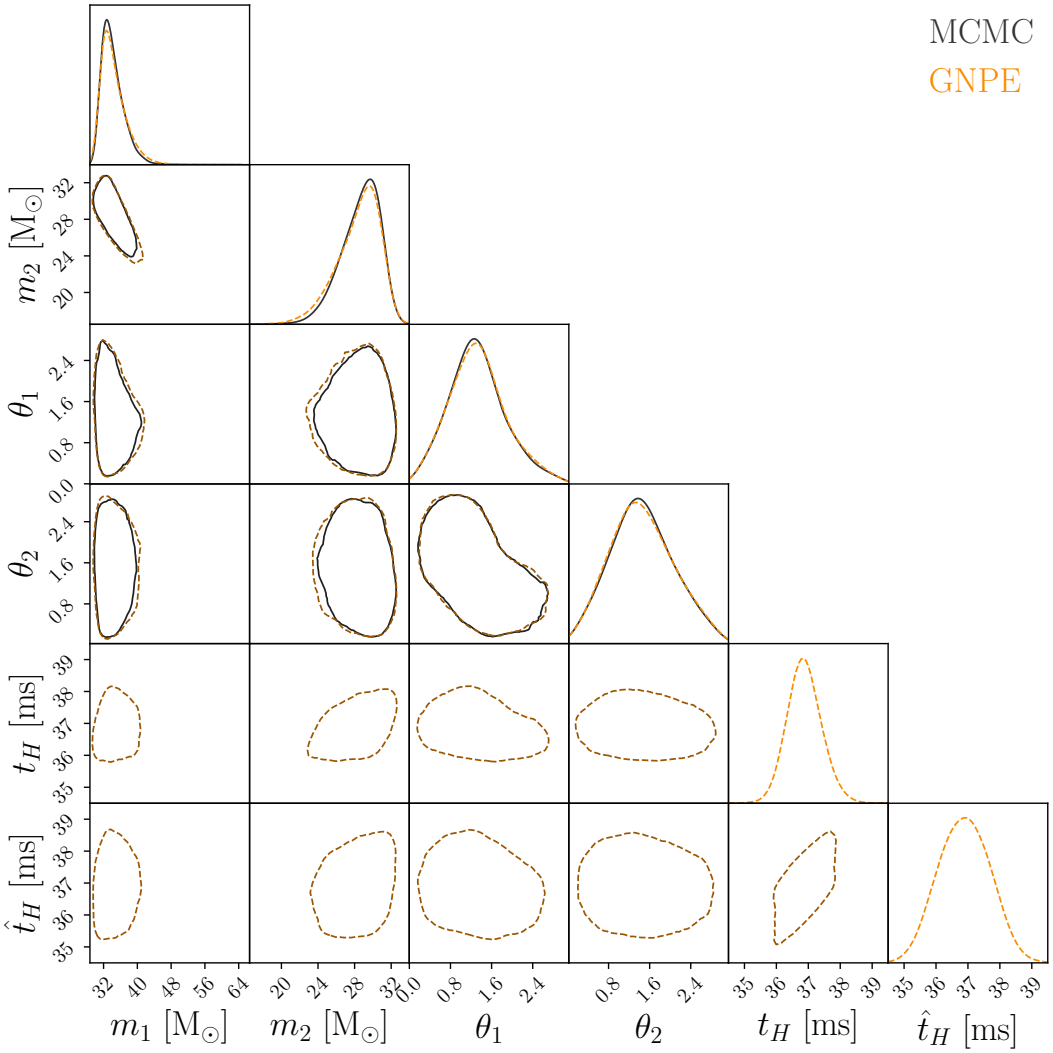

Figure C.5: Corner plot for the GW event GW170814 plotting 1D marginals on the diagonal and 90% credible regions for the 2D correlations. We display the two black hole masses m_1 and m_2 and two spin parameters θ_1 and θ_2 that are also shown in Fig. 4. We additionally display one of the pose parameters t_H and the corresponding proxy $\hat{t}_H$ from the last GNPE iteration. In training, the neural density estimator learned that the true pose t_H differs by at most 1 ms from the proxy $\hat{t}_H$ that it is conditioned on (since we chose a kernel $\kappa_{\text{narrow}} = U[-1 \text{ ms}, 1 \text{ ms}]^{n_I}$, see section 5.2). This explains the strong correlation between t_H and $\hat{t}_H$ we observe. For the same reason, the observed correlations between the $\hat{t}_H$ and the non-pose parameters $(m_1, m_2, \theta_1, \theta_2)$ are similar to those between the true pose t_H and $(m_1, m_2, \theta_1, \theta_2)$.

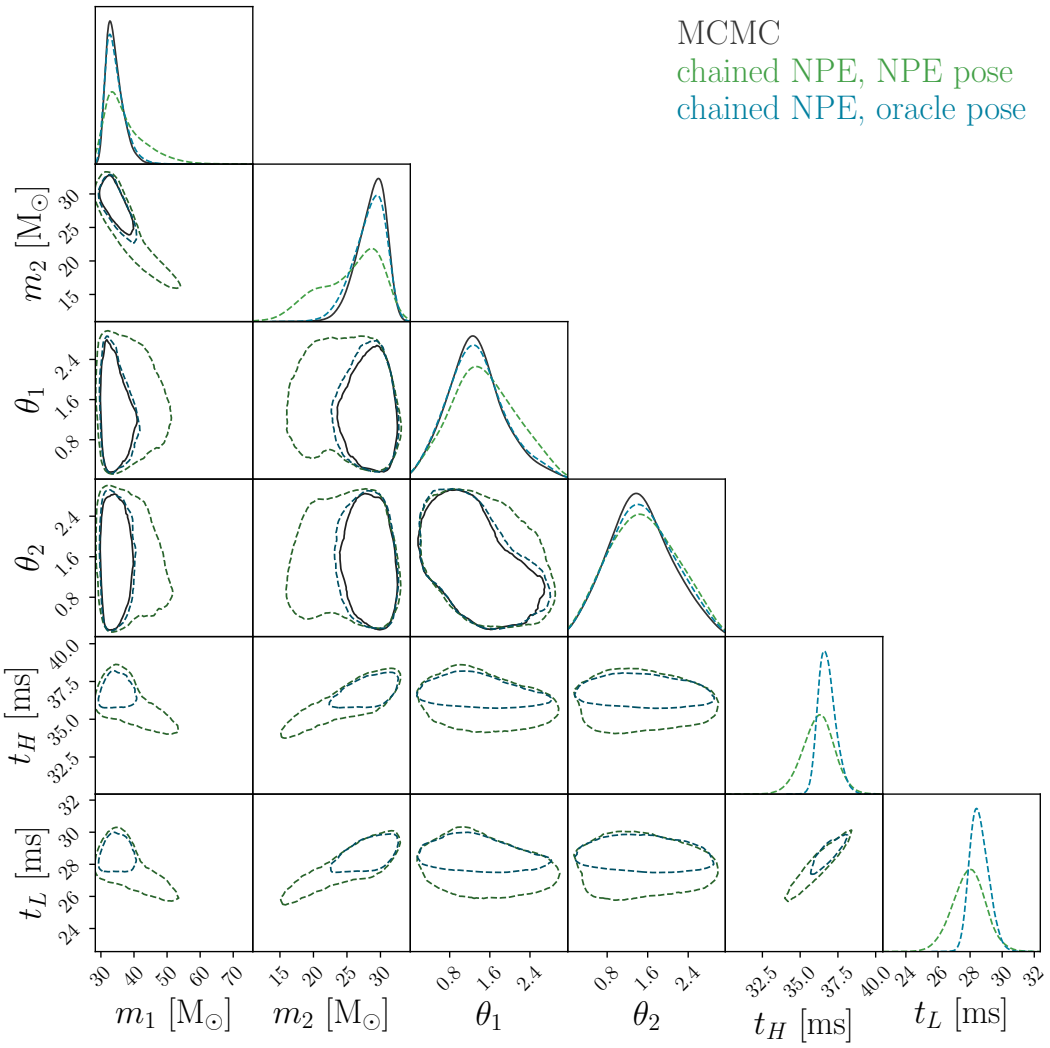

Figure C.6: Corner plot for GW170814 with the four parameters $(m_1, m_2, \theta_1, \theta_2)$ that are also displayed in Fig. 4, as well as the pose (t_H, t_L). We compare chained NPE as described in section 5.3 to an oracle version: for the earlier the pose is inferred using standard NPE (green) while for the latter we take an oracle pose provided by a (slow) nested sampling algorithm (teal). We observe, that the result using the oracle pose matches the MCMC reference posterior well, while the other one shows clear deviations. Both versions use the same density estimator for the non-pose parameters $\phi \subset \theta$. This demonstrates that inaccuracies of the chained NPE baselines can be almost entirely attributed to inaccurate inital estimates of the pose. Poor pose estimates can occur since the density estimator trained to extract the pose operates on non pose-standardized data.
Note: The MCMC reference algorithm LALInference does not provide full pose information since it automatically marginalizes over t_c. For the oracle pose we thus employ the nested sampling algorithm bilby (Ashton et al., 2019).

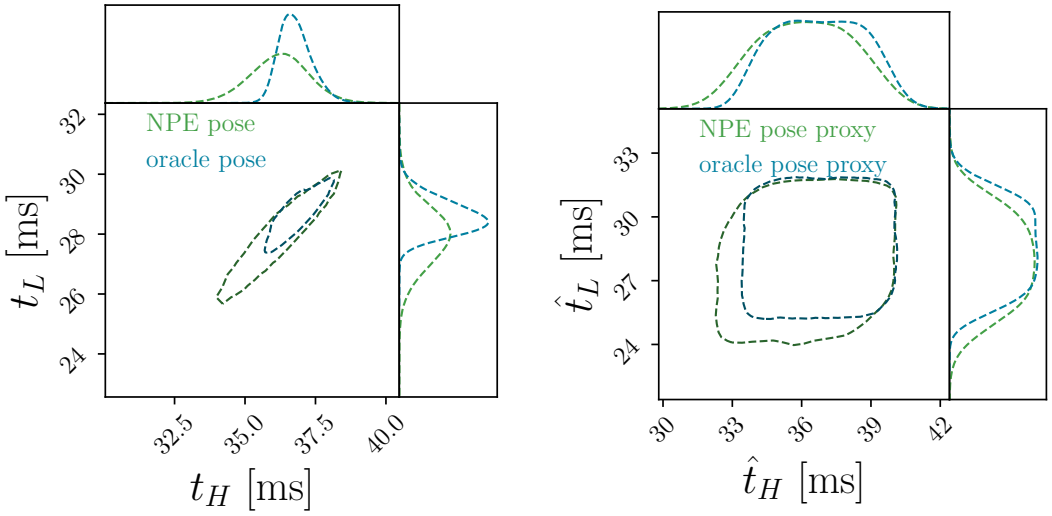

Figure C.7: Left: Pose parameters t_H and t_L for the GW event GW170814, estimated with the neural density estimator q_{init} with standard NPE (green), as well as the "true" pose inferred with bilby (teal). Right: Pose proxies $\hat{t}_H$ and $\hat{t}_L$ for the wide kernel $\kappa_{\text{wide}} = U[-3\text{ ms}, 3\text{ ms}]^{n_I}$. These are obtained from the pose estimates in the left panel via a convolution with κ_{wide}. We observe that the deviation between the oracle and the NPE estimate is substantially smaller for the pose proxy than for the pose itself due to the blurring operation. This leads to a better performance of fast-mode GNPE (with κ_{wide} and only one iteration) compared to chained NPE in section 5.3.