Gaussian Processes for Glitch-robust Gravitational-wave Astronomy

Gregory Ashton, 1★
Department of Physics, Royal Holloway, University of London, TW20 0EX, United Kingdom

7 March 2023

ABSTRACT
Interferometric gravitational-wave observatories have opened a new era in astronomy. The rich data produced by an international network enables detailed analysis of the curved space-time around black holes. With nearly one hundred signals observed so far and thousands expected in the next decade, their population properties enable insights into stellar evolution and the expansion of our Universe. However, the detectors are afflicted by transient noise artefacts known as “glitches” which contaminate the signals and bias inferences. Of the 90 signals detected to date, 18 were contaminated by glitches. This feasibility study explores a new approach to transient gravitational-wave data analysis using Gaussian processes, which model the underlying physics of the glitch-generating mechanism rather than the explicit realisation of the glitch itself. We demonstrate that if the Gaussian process kernel function can adequately model the glitch morphology, we can recover the parameters of simulated signals. Moreover, we find that the Gaussian processes kernels used in this work are well-suited to modelling long-duration glitches which are most challenging for existing glitch-mitigation approaches. Finally, we show how the time-domain nature of our approach enables a new class of time-domain tests of General Relativity, performing a re-analysis of the inspiral-merger-ringdown test on the first observed binary black hole merger. Our investigation demonstrates the feasibility of the Gaussian processes as an alternative to the traditional framework but does not yet establish them as a replacement. Therefore, we conclude with an outlook on the steps needed to realise the full potential of the Gaussian process approach.

Key words: gravitational-waves – black hole physics

1 INTRODUCTION
The emerging field of gravitational-wave astronomy is built on the back of a multi-decade effort to design and construct kilometre-scale interferometers that can measure their relative arm lengths to better than 1 part in 10^21. An international network of such detectors (Advanced LIGO, Virgo, and KAGRA (Aasi et al. 2015; Acernese et al. 2015; Akutsu et al. 2021)) are now in operation and have so far observed 90 compact binary coalescences (CBC) gravitational-wave signals, including binary black hole, neutron star-black hole, and binary neutron star mergers (The LIGO Scientific Collaboration et al. 2021a). However, of these detections, 18 are contaminated by transient non-Gaussian detector artefacts known as glitches. If left unaddressed, these contaminating noise sources bias astrophysical inferences, undermining the scientific outputs from this exquisite data (Powell 2018; Macas et al., 2022). In this work, we develop a novel solution to analyse signals contaminated by glitches, simultaneously modelling the glitch using a Gaussian process (Rasmussen 2004).

Glitches represent particular epochs in which a single detector misbehaves, often due to local environmental disturbances. Therefore, it is generally assumed that any glitch-generating process is independent between distinct detectors. This independence is essential for providing a means to distinguish them from astrophysical signals, which produce coherent power at multiple detector sites.

The characterisation of glitches forms an integral part of analysing data from interferometric detectors for two reasons (for a review, see Davis & Walker (2022)). First, studies which can identify the source of the glitch-generating process (e.g. by temporally correlating the signal as seen in the detector strain h(t) with auxiliary channels monitoring the instrument) provide a means to improve the quality of data directly resulting in a reduction in the number of glitches in future data. Such a reduction can naturally lead to an improvement in the ability of search algorithms to identify signals. The reduced background noise rate leads to more confidence in identifying true astrophysical signals. Second, and of relevance to this work, if a glitch overlaps a signal, care must be taken to carefully unpick the signal from the non-Gaussian noise (the glitch).

Coincidences between glitches and signals are a common problem, perhaps most distinctly demonstrated by the LIGO Livingston glitch which overlapped the first-observed binary neutron star merger GW170817 (Abbott et al. 2017; Pankow et al. 2018). Two approaches were applied to the mitigation of this glitch. A simple window function was applied for search analyses to zero out data around the glitch. Meanwhile, for source parameter estimation, the glitch was modelled by BayesWave (Cornish & Littenberg 2015), a flexible time-frequency wavelet reconstruction, and then a realisation of that model was subtracted from the data. This latter method, known colloquially as “deglitching”, has now become the de-facto standard approach to perform parameter estimation for the 20% of signals contaminated by glitches. While BayesWave remains the primary tool to model the glitch, a new tool gwsubtract (Davis et al. 2019) has also been used, which applies linear subtraction building the glitch model from a witness channel (Allen et al. 1999).

Davis et al. (2022) describes the basic deglitching process used to
analyse glitch-contaminated signals during the third and most recent observing run of the LIGO, Virgo, and KAGRA detectors. First, a check is performed for all candidate signals to see if they admit excess glitch power. Second, the glitch is modelled, and a realisation of the glitch drawn from the model is subtracted from the data. Finally, the deglitched data is analysed using a standard parameter estimation framework.

Deglitching has been highly successful in dealing with contaminated signals; the most recent study (Hourihane et al. 2022) demonstrates its general-purpose applicability. However, some drawbacks motivate us to study alternative approaches. First, the underlying algorithm that models the glitch is designed to study short-duration transient bursts. This makes it ideally suited to deglitching short-duration (≤1 s) glitches. However, Hourihane et al. (2022) find that long-duration glitches, such as the fast and slow-scattering glitches (discussed further below), are challenging and need specialised settings. The need for specialised settings calls for manual analyses: a slow process requiring an expert to repeatedly analyse an event and, often by eye, decide if the glitch has been adequately excised. Second, the principle of deglitching is inherently flawed because it ignores uncertainty in the glitch model. To deglitch the data, some draw from this model is taken. Usually, this is either the median or a random draw. However, Hourihane et al. (2022) find that there are differences in the deglitched data depending on which draw is taken for slow scattering glitches. Such differences indicate that the glitch model has a non-negligible uncertainty which is not correctly accounted for when estimating the parameters of the astrophysical signal from the resultant deglitched data. The solution to this problem is to analyse the signal and glitch together. Already, Hourihane et al. (2022) have demonstrated how the underlying deglitching algorithm can be extended to this end.

Proper glitch mitigation is vital for ensuring unbiased astrophysical inferences of the signal. Two clear examples of the difficulties inherent in the current process came to light during the third observing run. The first example is the event GW200129, a binary black hole system with the largest network signal-to-noise (SNR) ratio of any observing run. The first example is the event GW200129, a binary black hole system with the largest network signal-to-noise (SNR) ratio of any observing run. The first example is the event GW200129, a binary black hole system with the largest network signal-to-noise (SNR) ratio of any observing run. The first example is the event GW200129, a binary black hole system with the largest network signal-to-noise (SNR) ratio of any observing run. The first example is the event GW200129, a binary black hole system with the largest network signal-to-noise (SNR) ratio of any observing run. The first example is the event GW200129, a binary black hole system with the largest network signal-to-noise (SNR) ratio of any observing run. The first example is the event GW200129, a binary black hole system with the largest network signal-to-noise (SNR) ratio of any observing run. The first example is the event GW200129, a binary black hole system with the largest network signal-to-noise (SNR) ratio of any observing run. The first example is the event GW200129, a binary black hole system with the largest network signal-to-noise (SNR) ratio of any observing run. The first example is the event GW200129, a binary black hole system with the largest network signal-to-noise (SNR) ratio of any observing run. The first example is the event GW200129, a binary black hole system with the largest network signal-to-noise (SNR) ratio of any observing run. The first example is the event GW200129, a binary black hole system with the largest network signal-to-noise (SNR) ratio of any observing run.
\[ f_s, \text{ we take the discrete Fourier transform (DFT):} \]
\[ \hat{d}_j^{(f)} = \frac{1}{f_s} \sum_k \hat{d}_k^{(f)} e^{-2\pi i j k / f_s T}, \]  
\[ (2) \]

where \( j \) indexes the frequency bin and a tilde denotes a frequency-domain quantity. Then, assuming the background noise and the signal are additive and that the noise is stationary and zero-mean with a known PSD \( P(f) \), the single-detector log-likelihood is given by the Whittle likelihood:

\[ \ln L(d^{(f)}|\theta, M) = -\frac{1}{2} \sum_j \ln(2\pi P_j) - \frac{2}{T} \sum_j \frac{[\hat{d}_j^{(f)} - \hat{\mu}_j(\theta)]^2}{P_j}. \]  
\[ (3) \]

where \( j \) runs over the positive frequency bins and \( \hat{\mu}_j(\theta) \) is the frequency-domain representation of the signal model \( M \) for parameters \( \theta \). While common in the general field of time-series data analysis, the term Whittle likelihood is rarely used in the gravitational-wave literature (see, e.g. Veitch et al. (2015)). We learned of the name from Thrane & Talbot (2019) and chose to apply it here to distinguish it from the GP likelihood, which we derive later in this text. If data from multiple detectors are available, assuming that the noise is uncorrelated, the network log-likelihood is constructed from the sum of the log-likelihood for each detector (and this applies to both the Whittle likelihood and GP approaches).

Eq. (3) was first derived by Peter Whittle (Whittle 1953) as an approximation to the Gaussian likelihood for a stationary time series:

\[ \ln L(d|\theta, M) = -\frac{1}{2} \ln \Sigma^{-1} r(\theta) - \frac{1}{2} \ln \left( (2\pi)^N |\Sigma| \right), \]  
\[ (4) \]

where \( r(\theta) = d - \mu(t, \theta) \) is the residual after subtracting the signal from the data, \( \Sigma \) is the covariance matrix of the background noise, and \( N = f_s T \) is the number of data points. The Whittle likelihood approximation (Eq. (3)) is advantageous compared to the full Gaussian likelihood (Eq. (4)) because the DFT can be replaced by the fast Fourier transform (FFT) algorithm, reducing the computational cost from \( \mathcal{O}(N^3) \) to \( \mathcal{O}(N \log N) \). A recent pedagogical review (Rao & Yang 2021) demonstrated the differences between the two likelihoods by deriving a frequency-domain representation of the Gaussian likelihood and found the difference to be of order \( \mathcal{O}(1/N) \). For typical gravitational-wave inference problems, a sampling frequency of at least a few kHz is needed to ensure the Nyquist frequency captures the signal properties during the merger and ringdown, and a duration of a few to tens of seconds is typical. Therefore, we expect only slight differences between the Whittle and Gaussian likelihoods. However, there are known deficiencies in the Whittle likelihood approach. For example, recently, Talbot & Thrane (2020) and Talbot et al. (2021) have investigated the impact of the fixed PSDs and windowing of the time series, respectively. Each of these could bias individual results at a level equivalent to other systematic uncertainties, and the effect can be magnified in population studies.

For the prior distribution of Eq. (1), as a general rule, the standard approach is to apply uninformative priors where possible (e.g. isotropic priors on the sky position). However, in some cases, these are informed by the output of the detection pipeline (e.g. we restrict the time of the signal to some relevant range) or by other astrophysical data (e.g. the observation of an electromagnetic counterpart may imply certain restrictions on the orientation of the source). In this work, we apply standard prior definitions used throughout the field (see, e.g. Romero-Shaw et al. (2020)).

Having defined the likelihood and priors, to estimate Eq. (1), either a stochastic-sampling (Veitch et al. 2015; Cornish & Littenberg 2015; Biwer et al. 2019; Ashton et al. 2019) or grid-based (Lange et al. 2018) algorithms is used. These methods are computational-intensive, typically needing \( \mathcal{O}(10^8) \) or more evaluations of the likelihood and prior, with each evaluation taking \( \geq 1 \times 10^{-3} \) s (typically, this time is dominated by the generation of the waveform and the overheads of the Whittle likelihood). Therefore, the typical time needed to estimate the posterior distribution is about a day. However, this cost can increase dramatically if, for example, more physically accurate waveforms are used or the signal duration requires a longer data span. The total wall time required for analysis can be reduced by parallelisation and optimisation of the likelihood in addition to methods that reduce the cost of the likelihood and waveform evaluation. Nevertheless, Bayesian inference remains expensive but necessary to resolve the highly-correlated structure in the posterior distribution robustly (whereby “robustly” we mean that repeated analyses reproduce the same result within statistical uncertainties).

Eq. (3) has been applied with great success to every CBC signal observed by advanced-era detections (The LIGO Scientific Collaboration et al. 2021a). However, it cannot be applied directly because, as discussed in the introduction, up to 20% of these are contaminated by glitches, violating the fundamental assumption that the background noise is stationary and Gaussian. To address this, the BayesWave algorithm extends Eq. (1), including an additional glitch model (and associated set of model parameters). Performing inference on this glitch model (simultaneously with the signal model), a model of the glitch is constructed and then subtracted from the data. The standard inference process then uses the glitch-subtracted data in Eq. (3). The difficulty arises if the glitch is imperfectly removed. Then, the residual power will remain in the glitch-subtracted data and bias the inference of the signal properties. In the next section, we describe our proposed modification to the standard methodology, replacing Eq. (3) with a GP likelihood.

### 2.2 Gaussian processes

A GP time series as a multivariate Gaussian distribution with a mean function \( \mu(t; \theta) \) and a kernel function \( k(t_m, t_n; \alpha) \) where \( \alpha \) is a set of hyperparameters describing the noise process. Under this interpretation, the log-likelihood is given by

\[ \ln L(d|\theta, \alpha, M) = -\frac{1}{2} \ln \left( 2\pi \Sigma(\alpha)^{-1} r(\theta) - \frac{1}{2} \ln |\Sigma(\alpha)| \right). \]  
\[ (5) \]

Comparing the GP likelihood with Eq. (4), we see that we have replaced the (fixed) covariance matrix \( \Sigma \) with the covariance matrix generated by the kernel function:

\[ \Sigma_{mn}(\alpha) = k(t_m, t_n; \alpha). \]  
\[ (6) \]

GPs are powerful because they enable us to model the noise process instead of the noise realisation. The art of GP modelling is constructing a kernel that captures the noise process and inferring the hyper-parameters. Simultaneously, we can infer the properties of the mean model \( \mu(t; \theta) \), which will be marginalised with respect to the uncertainty in the GP noise model.

Direct application of Eq. (5) is infeasible for gravitational-wave signals since, as with the full Gaussian likelihood the computation cost is of order \( \mathcal{O}(N^3) \). However, if the kernel function is stationary and constructed from a mixture of exponential functions, Foreman-Mackey et al. (2017) demonstrated that the likelihood can be calculated with linear scaling by exploiting structure in the covariance matrix. The authors implement this method in the celerite software, which we will use throughout this work. Such kernel functions are not abstract: they arise naturally from random processes consisting
of stochastically-driven damped harmonic oscillators. In a meaningful sense, many of the noise processes afflicting gravitational-wave detectors can be considered a collection of harmonic oscillators. Therefore, it is natural to hypothesise that such fast and scalable GPs may be brought to bear on the problem of gravitational-wave data analysis. To our knowledge, GPs have yet been applied to this problem (though they have been applied in other settings in the field, for example to estimate model uncertainty (Moore et al. 2016; Doctor et al. 2017; Williams et al. 2020) and for interpolating posterior inferences (D’Emilio et al. 2021)).

2.3 The GP likelihood for gravitational-wave signals

In this work, we will utilise the celerite software to develop a gravitational-wave GP likelihood capable of inferring the properties of gravitational-wave signals contaminated by glitches. We will exclusively use the Simple Harmonic Oscillator (SHO) kernel described in Foreman-Mackey et al. (2017), which is characterised by an oscillator frequency $\omega_0$, a quality factor $Q$, and amplitude $S_0$. The SHO kernel is stationary: it depends only on the difference $\tau_m - \tau_n$ (this is true of all kernels in celerite). Therefore, the stationary SHO kernel would appear to be a poor fit to model non-stationary transient glitches. Indeed, in practice, the SHO kernel performs poorly for short-duration glitches that last for only a fraction of a second. However, as we demonstrate later, the SHO kernel can model the underlying noise process for glitches that last an appreciable fraction of the signal duration. We discuss this further in Section 4 and suggest new avenues for applying transient GP models.

In principle, a GP model could be built that models the coloured Gaussian background noise characteristic of gravitational-wave detectors. This could be done using a mixture of SHO terms using terms with small quality factors to model the coloured noise and SHO terms with large quality factors to model the narrow Lorentzian features (see Littenberg & Cornish (2015) for a discussion of how BayesWave performs a similar task). However, for this initial study, where we are primarily focused on the ability of the GP to model glitches, we will instead utilise a technique known as pre-whitening.

To pre-whiten the data, we first construct a PSD $P_f$ using the median Welch method (Welch 1967). We then FFT the time series data to obtain $\hat{d}$ and then apply the whitening transform:

$$d_i = \text{IFFT} \left( \frac{2}{N} \frac{d}{\sqrt{P}} \right),$$

where IFFT is the inverse FFT and the factor of $2/N$ normalises the whitened data. If the PSD properly represents the coloured Gaussian noise, then the whitened data $\hat{d}$ is zero mean with unit variance.

Our GP model is then applied to whitened data $\hat{d}$ using a fixed PSD. To calculate the likelihood given a set of GP hyperparameters $\alpha$ and model parameters $\theta$, we apply Eq. (5) but note that the mean model $\mu(\tau; \theta)$ must also be whitened by the same PSD.

By default, our GP kernel (Eq. (6)) then consists of a white-noise term with a single hyperparameter $\sigma$ and model parameters $\theta$; we apply Eq. (5) but note that the mean model $\mu(\tau; \theta)$ must also be whitened by the same PSD.

We take a representative example identified by GravitySpy (Zevin et al. 2017; Glanzer et al. 2022) in the LIGO Livingston interferometer during the O3 observing run. In the left-hand panel of Fig. 1, we plot a spectrogram of the data illustrating three distinct scattering arches between 10 and 70 Hz. Overlaid on this figure is the time-frequency track of the simulated signal that we will add to the data (it is not included in the spectrogram to avoid confusion with the glitch itself). Before adding the simulated signal, we apply the BayesWave deglitching routine using standard settings. The resulting spectrogram (right-hand panel of Fig. 1) demonstrates some residual glitch power below 20 Hz which we were unable to remove, but this residual power does not fall along the frequency-time track of the injected signal.

In contrast to the analyses of overlapping signals and glitches in real data, the deglitching process in this case study is applied in the absence of a signal. Therefore, it is not representative of a real analysis but does represent an ideal of the deglitching process: the glitch can be characterised and subtracted without signal confusion. However, in practice, this will not usually be possible unless an auxiliary channel provides a means to model the glitch without using the strain data. Instead, for typical cases, which use BayesWave to deglitching the data, the signal and glitch must be modelled simultaneously. This leaves open the possibility for erroneous signal removal. Nevertheless, it is a good demonstration of how the GP method compares against the ideal deglitching scenario.

Having both the original data and the deglitched data, we can perform a direct comparison of three different approaches. The first, which we label “BW+WL” applies the standard Whittle likelihood to data which has been deglitched using BayesWave. Therefore, BW+WL denotes the method used by the LVK collaborations to
date. The second, which we label “GP”, is to apply the GP method directly to the original data, modelling the glitch and signal simultaneously. Finally, we label analyses where the Whittle likelihood is applied directly to the original data “WL”; these analyses illustrate the extent to which the glitch biases inferences in the absence of a mitigation technique. In validating the method, we did also analyse the deglitched data using the GP approach. As expected, we found near identical results to the BW+WL approach since the GP simply sets the hyperparameters of the SHO such that its effect is negligible, and then one recovers the standard Whittle likelihood.

In each analysis, taking either the original or deglitched data, we add a simulated equal-mass non-spinning BBH signal generated using the time-domain IMRPhenomT waveform approximant (Estelles et al. 2022). The process of generating this analysis (i.e. obtaining the strain data and adding the simulated signal) is independent of the analysis method, ensuring a straightforward comparison can be made. The signal track of the simulated signal is shown as a dashed curve in Fig. 1.

We then analyse the original data (including the simulated signal) using the WL and GP approach and the deglitched data using the BW+WL approach. All analyses are performed on data from a single detector, though they naturally scale to multiple detectors by adding together the log-likelihoods assuming independent noise. For the comparative case studies (A and B), we infer the signal properties using the IMRPhenomT waveform, which assumes the spin of both black holes is aligned along their angular momentum and we a priori fix the sky position of the source to the simulated values. We apply these simplifications because they reduce the overall computational cost by nearly an order of magnitude with minimal impact on the comparison; ultimately, we are interested in the ability of the GP model to separate the signal from the noise. However, we do note that in this feasibility study, we have therefore not examined the interaction of a signal with misaligned spins, which will exhibit precession effects, with the GP method. Finally, for all analyses, we generate a PSD by applying the median Welch method to off-source data. For the WL and BW+WL analysis, the PSD is used directly in the likelihood, while for the GP method, the PSD is used to whiten the data and the predicted model.

In Fig. 2, we compare kernel density estimates of four source parameters inferred using the WL, GP, and BW+WL approaches. We choose to compare the detector-frame\(^1\) chirp mass (Finn & Chernoff 1993):

\[
M^d = \frac{(m_1^d m_2^d)^{3/5}}{(m_1^d + m_2^d)^{1/5}},
\]

(where \(m_1^d\) and \(m_2^d\) are the detector-frame mass of the two black holes), the mass ratio:

\[
q = m_2^d/m_1^d,
\]

the effective spin (Santamaria et al. 2010; Ajith & Bose 2009):

\[
\chi_{\text{eff}} = \frac{\chi_1 + q\chi_2}{1 + q},
\]

where \(\chi_1\) and \(\chi_2\) are the aligned spin components of the two black holes, and the luminosity distance, \(d_L\).

\(^1\) For gravitational-wave radiation from sources at cosmological distances, the signals are red-shifted. As a result, it is not possible to measure the source-frame mass directly, but instead, we measure the detector-frame mass. The former can be inferred by supplying a cosmological model to estimate the redshift \(z\) from the measured luminosity distance and then scaling the detector-frame mass by \(1/(1 + z)\) (Cutler & Flanagan 1994).

The kernel used in the GP analysis presented in Fig. 2 is formed from the addition of a white noise component (characterised by its standard deviation) and \(M = 3\) SHO terms (characterised by the frequency, quality and amplitude factors) resulting in a total of 10 noise hyperparameters. The choice of three SHO terms and the prior applied to their hyperparameters was reached by manual tuning; using a greater number of SHO terms, their effect was negligible, while using fewer terms, the source-parameter inferences were biased. In this instance, the optimal number of three was also suggested by the observation of the spectrogram Fig. 1 where we observe three distinct features. It is then unsurprising that the median SHO frequency inferred from the GP analysis coincides nicely with the three arches in Fig. 1 (see horizontal dashed lines). But, this also illustrates a clear deficiency of the GP model: we assume the GP kernel is stationary, but Fig. 1 suggests that the frequency is, in fact, changing with time. Moreover, models of light scattering (Longo et al. 2020) have previously demonstrated the frequency is changing with time. Therefore, we predict that non-stationary kernels may produce better results for transient glitches with a short duration relative to the length of data being studied. In future, we could investigate the use of the Bayesian evidence to select the optimal number or a Reversible-Jump MCMC able to optimize over the number. Finally, we add that for all three SHO terms, the median inferred quality factor was ~ 30.

Let us now consider each of the analyses in Fig. 2 in turn.

First, for the WL analysis, which applies no glitch mitigation, the inferred source parameters are severely biased: this is clearly demonstrated by the measured chirp mass and mass ratio where the peak of the distribution lies well away from the simulation value. This fact is well known (see, e.g., (Pankow et al. 2018)). In recent work, we demonstrated (Ashton et al. 2022) that analysing scattering glitches alone with a BBH model resulted in inferred mass ratios of ~ 0.2. That the posterior for the mass ratio under the WL analysis in Fig. 1 peaks around this value is therefore consistent with the interpretation that the WL analysis effectively *misses* the simulated signal and tries to fit the BBH model to the glitch instead.

Second, the BW+WL approach is successful: all four parameters are recovered within the 90\% interval. This observation is entirely unsurprising, of course: the success of the BW+WL approach is already well known in the literature, and Fig. 1 demonstrates visually that the BayesWave algorithm successfully deglitches the data.

Finally, like the BW+WL approach, the GP approach also successfully recovers the simulated values within the 90\% credible intervals. This indicates it has successfully modelled the signal as part of the mean model and the glitch as part of the noise. To further check this, in Fig. 3, we create a posterior predictive test showing the modelled glitch and noise alongside the simulated signal. The posterior predictive plot demonstrates the difficult task the GP analysis has completed: to pick out the signal from the highly non-Gaussian noise.

Comparing the GP and BW+WL approaches in Fig. 2, we see that the GP credible interval is, in all cases, wider than that of the BW+WL method. Taken at face value, this suggests the BW+WL method is preferable since it constrains the source parameter with greater precision. However, we remind the reader that the BW+WL approach is operating under perfect conditions, where the glitch is modelled in the absence of the signal. In practise, this will not be the case. On the other hand, the GP approach results are marginalised over the uncertainty induced by the glitch, which increases the uncertainty. Therefore, the increase in uncertainty represents exactly the goal of the GP analysis: to jointly constrain both the astrophysical signal and glitch and circumvent the need for deglitching which intrinsically neglects uncertainty. However, we do also remark that
**Figure 1.** Spectrograms of the slow-scattering glitch analysed in case study A plotted alongside the frequency track of the simulated signal. In the left-hand panel, we show the original strain data, whereas the right-hand figure shows the strain data after BayesWave glitch subtraction is applied. Note that the glitch subtraction is applied before the simulated signal is added. We also show dashed horizontal lines in the left-hand panel highlighting the median SHO frequency inferred in the GP analysis.

**Figure 2.** The posterior distribution for the detector-frame chirp mass $M^d$, mass ratio $q$, effective spin $\chi_{\text{eff}}$, and luminosity distance $d_L$ for case study A. In each figure, we compare the WL applied to the original data (WL), the GP applied to the original data (GP), and the WL applied to the deglitched data (BW+WL). The simulated signal value is given as vertical dashed lines while 90% credible intervals are given as coloured vertical lines for each distribution.
the BW+WL and GP glitch models are fundamentally different and, it could be the case that the BayesWave model is in some way better than the SHO model. To test this, we will need to compare a full joint analysis by BayesWave of the glitch and BBH signal as done in Hourihane et al. (2022). Hence, for the time being, we can conclude that the GP process is able to model the glitch and signal simultaneously, but further comparisons are needed to understand the relative performance of the two approaches.

### 3.2 Case Study B: An O3 fast-scattering glitch

In our second case study, again using single-detector data, we add a simulated signal to data containing a fast-scattering glitch. The original data is shown in Fig. 4 and illustrates the typical behaviour of this glitch class: a series of short-duration artefacts centred around (in this instance) a frequency of ~60 Hz. We did try to excise the glitch from the original data. However, we found it difficult to find adequate BayesWave configuration which removed the glitch (though Hourihane et al. (2022) do report successful modelling of other fast scattering glitches). For this reason, we cannot perform the BW+WL analysis. Therefore, we instead perform a repeated WL-offset analysis but shift the simulated signal time by +20 s. During this time, the data does not contain any transient noise artefacts (though we note the estimated PSD will).

To the original data, we add a simulated signal (identical to that of case study A); the time-frequency evolution of this signal is shown in Fig. 4 and then perform three analyses: the WL, GP and WL-offset analysis. For the GP model, we find that a kernel consisting of a single SHO term in addition to a white-Gaussian noise term was sufficient to model the background noise and extract the signal properties. In Fig. 4, we plot the median inferred frequency of the SHO term as a horizontal bar demonstrating that it captures the average frequency of the repeating transient artefacts. The quality factor the SHO term inferred from the data was ~10.

In Fig. 5, we plot kernel density estimates of the posterior from the three analyses. Once again, the WL analysis demonstrates the severe bias by an analysis which is unable to model the glitch. Meanwhile, the GP and WL-offset analyses both recover the signal within their 90% credible intervals. However, the WL-offset analysis is better constrained. This is to be expected for the same reasons discussed in the context of case study A. The difference is even more severe in this case. We expect that this is due to the particularity of the glitch under study.

In Fig. 6, we create a posterior-predictive plot which demonstrates the overlap of the signal and glitch in the inspiral phase. From this plot, we can understand why the WL analysis, which assumes only Gaussian noise and a BBH signal, is so severely biased by the oscillatory features of the glitch. Furthermore, we can understand why the WL-offset analysis is far better constrained than the GP analysis. For the GP analysis, the inspiral is severely biased by the glitch, effectively reducing the SNR and hence resulting in a wider posterior compared to the WL-offset analysis where the inspiral is uncontaminated by the glitch.

---

**Figure 3.** A posterior predictive plot for the slow-scattering glitch (case study A). In a solid grey curve, we plot the whitened strain data from LIGO Livingston; this includes the simulated signal added to the data, which we also plot as a solid black line (after whitening). In purple, we plot the posterior model prediction: a solid curve is generated from a median estimate of the source parameters, while the purple-shaded region indicates the 90% credible interval for the model prediction (generated by drawing repeated predictions from the model from the posterior distribution). Finally, in a solid grey band, we plot the 90% credible interval for the inferred GP noise by repeatedly drawing predictions from the GP kernel with a zero mean. The upper panel provides a close look at the features of the signal near to the merger, the lower panel includes a wider span of data illustrating the features of the glitch (cf. Fig. 1).

**Figure 4.** Spectrogram of the fast-scattering glitch analysed in case study B plotted alongside the frequency track of the simulated signal. We also show dashed horizontal lines highlighting the median SHO frequency inferred in the GP analysis.
Figure 5. The posterior distribution for the detector-frame chirp mass $M_d$, mass ratio $q$, effective spin $\chi_{\text{eff}}$, and luminosity distance $d_L$ for case study B. In each figure, we compare the WL applied to the original data (WL), the GP applied to the original data (GP), and the WL applied to an offset data segment without a glitch (WL-offset). The simulated signal value is given as vertical dashed lines while 90% credible intervals are given as coloured vertical lines for each distribution.

Figure 6. A posterior predictive plot for the fast-scattering glitch (case study B). In a solid grey curve, we plot the whitened strain data from LIGO Livingston; this includes the simulated signal added to the data, which we also plot as a solid black line (after whitening). In purple, we plot the posterior model prediction: a solid curve is generated from a median estimate of the source parameters, while the purple-shaded region indicates the 90% credible interval for the model prediction (generated by drawing repeated predictions from the model from the posterior distribution). Finally, in a solid grey band, we plot the 90% credible interval for the inferred GP noise by repeatedly drawing predictions from the GP kernel with a zero mean.

3.3 Case Study C: Enabling time-domains tests of General Relativity: Inspiral-Merger-Ringdown

In our final case study, we discuss how the GP approach can enable time-domain tests of General Relativity (GR) which are robust to glitches. The violent final stages of a binary black hole collision offer a unique opportunity to probe the limits of GR in the strong-field regime (see The LIGO Scientific Collaboration et al. (2021c) for the latest results). There are a variety of proposed tests, including theory-independent consistency tests (looking for generic fails of the predicted GR model) and tests which search for explicit predictions (see, e.g. Yunes & Siemens (2013); Yagi & Stein (2016) for reviews). However, most of these tests rely on the frequency-domain data analysis approach developed for source parameter estimation. While this is not inherently problematic, there is a subset of cases where the frequency-domain approach forces awkward workarounds to implement what are intrinsically time-domain tests.

An example of this is the so-called Inspiral-Merger-Ringdown (IMR) consistency test which, conceptually, breaks the data in the time domain into two segments, the inspiral and merger-ringdown (Abbott et al. 2016b; Ghosh et al. 2016, 2018). The IMR test then checks for consistency between the mass and spin of the remnant Kerr black hole, as predicted by the inspiral and merger-ringdown data separately. The difficulty arises in that the analysis is performed in the frequency domain, not the time domain. The separation of inspiral and merger-ringdown data is practically performed by choosing a cutoff frequency. Data below the cutoff is assumed to arise
from the inspiral, while data above the cutoff is taken to be of the merger-ringdown. However, this practical choice can result in spectral leakage if the inspiral signal contains power at frequencies above the cutoff or the merger-ringdown signal contains power below the cutoff. Ghosh et al. (2018) demonstrate that if the cutoff frequency is chosen to be that of the innermost stable circular orbit (ISCO) of the final Kerr back hole, then the amount of leakage is small. But, we point out that this validation was performed on a circular aligned-spin system. For precessing systems (and systems including the effects of eccentricity), the frequency evolution during the inspiral can be non-monotonic, and the signal can contain frequency content above the ISCO, which would result in spectral leakage between the inspiral and merger-remnant. Spectral leakage is of concern because, if it occurs at a significant level, it violates the premise of the test and enforces an automatic consistency. Moreover, even if the level of spectral leakage is small, the practical implementation of the IMR consistency test (splitting the signal in the frequency domain) is at odds with the conceptual idea of splitting the signal in the time domain.

A second example where a similar difficulty is faced is that of the ringdown-only analysis (Abbott et al. 2016b; Isi et al. 2019; Carullo et al. 2019; Capano et al. 2021), where data arising from the ringdown alone is analysed for consistency with the theoretical predictions of GR. Such tests face a similar difficulty to the IMR tests: they need to take data only from the ringdown, ensuring there is no contamination from the inspiral. In this case, a simple frequency cut will not suffice. The two dominant approaches developed to solve the problem are “gating and in-painting” (Capano et al. 2021) where the inspiral is replaced such that the over-whitened data is 0 in the region of interest and a time-domain likelihood (Carullo et al. 2019; Isi et al. 2019) which is able to analyse only the merger-ringdown.

Time-domain ringdown analyses demonstrate the general-purpose applicability of time-domain approaches to testing GR. That our GP approach is itself time-domain by construction motivates us to demonstrate how it can be readily applied to resolving the spectral leakage problem for IMR tests. To this end, we perform a reanalysis of the first observed binary black hole merger, GW150914 (Abbott et al. 2016a). Taking open data from the Gravitational-Wave Open Science Centre (Abbott et al. 2021), we analyse the data from LIGO Hanford and Livingston using the precessing IMRPhenomP3 (Estéllés et al. 2021). We use standard prior configurations (isotropic priors on the spin, uniform in the component masses, and uninformative on the position and orientation), noting that unlike case studies A and B, this analysis makes no assumptions about the spin or sky location. We first construct a PSD using the median Welch method from 124 s of data prior to the signal. We then analyse the full signal (labelled IMR), comprising 4 s of data centred on the GPS trigger time \( t_0 = 1126259462.418 \), the inspiral signal comprising 2 s of data before \( t_0 \), and the merger-ringdown (labelled Post-Ins.) comprising 2 s of data after \( t_0 \). The posterior distributions for the final mass and spin inferred from these three analyses are then presented in the upper panel of Fig. 7. Next, we apply the procedure described in Ghosh et al. (2018), (see Eqn 5), to calculate the fractional difference between the inspiral and post-inspiral analyses normalized to their average and plot the resulting distribution in the lower panel of Fig. 7.

Our results agree with the original analyses (Abbott et al. 2016b); we find strong consistency between the inspiral and merger-ringdown predictions. Moreover, the prediction of GR falls comfortably within the 90% credible interval. However, we do note that our 90% credible interval for the fractional deviation is smaller than that of the original analysis and our post-inspiral inference is similarly better constrained. In absence of the differing methodology, there are already several possible causes for this, including the use of a different waveform model and PSD. However, we anticipate that the most significant differences arise due to the construction of the time-domain cutoff. Namely, we cut the data based on the trigger time \( t_0 \) calculated from a full IMR analysis. However, this value refers to the peak of the 2-2 mode as measured at the Earth’s centre. Meanwhile, the light-travel-time delay means that the transition between merger and ringdown (itself a fuzzy concept) occurs at different times for the data from the two detectors. This fact can clearly be seen in Fig. 8 where we plot the data and the posterior predictive whitened waveforms from the IMR, Inspiral, and Post-Ins analyses. Here, one can see that the time-domain cutoff between Inspiral and Post-Ins occurs earlier in Hanford (relative to the overall waveform morphology) than in Livingston. A straightforward improvement would therefore be to define a detector-dependent cutoff time. However, like with the frequency-domain analysis, the choice of cutoff is in some sense arbitrary and will invariably produce different results. But crucially, for our time-domain analyses, any choice of cutoff is still a valid test of GR since we cleanly separate the data without the risk of spectral leakage. Therefore, for this presentation, we choose to keep our simplified analysis fixing the cutoff time universally. In future work, we plan
to investigate detector-dependent cutoff times by choosing a fixed phase of the waveform and correcting for the light travel time delay. Another interesting possibility opened up by the time-domain nature of our approach is to extend the number of data segments beyond just two and study consistency between them. Such a multiple-segment approach is to extend the number of data segments beyond just two and study consistency between them. Such a multiple-segment approach to model the inspiral and merger.

Since GW150914 was not contaminated by glitches, we did not include any SHO terms in the GP kernel. In testing, we found that adding SHO terms did not change the results: we found near-identical inferences about the signal parameters, while the posterior distributions for the glitch SHO terms were uninformative except for setting a limit on the maximum amplitude. In a sense, adding additional terms in the GP kernel when fitting to a signal that is thought to be uncontaminated amounts to an automated residual test of GR (an analysis where the best-fit signal is subtracted from the data and the residual is searched for excess power). If the GP term were found to have non-zero power and one was certain the signal was not contaminated by a glitch, this would suggest that the signal model was not able to fit some features of the data. Such a result could suggest the presence of an unmodelled part of the signal.

It is worthwhile pointing out that, since we do not include any SHO term, this case study bears parallels with extensions of the time-domain ringdown analyses to also model the inspiral and merger. However, our new method can be easily extended to also include SHO terms to model any contaminating glitches: this could be crucial to future detections to verify that signal power is arising from an astrophysical source and not terrestrial contamination. In future work, we will explore how the additional advantages of the GP approach are leveraged. For example, using the GP to explicitly model the coloured Gaussian noise (so that pre-whitening is not needed) and performing time-domain GR tests on signals contaminated by glitches.

4 CONCLUSION AND DISCUSSION

We have presented a feasibility study for a novel approach to analysing transient signals observed by ground-based gravitational-wave detectors using Gaussian processes. This new method fundamentally differs from the traditional approach in that it operates in the time domain and is able to model the physical processes generating Gaussian and non-Gaussian noise alongside an astrophysical signal. Our first two case studies demonstrate that the true source parameters of a signal contaminated by scattered light glitches can be recovered. With 20% of observed signals observed by current generation ground-based detectors contaminated by signals, this marks an important step towards enabling joint analysis of glitches and signals and a ready alternative to explicit glitch modelling. Then, in our final case study, we demonstrate that the Gaussian process approach can be easily leveraged to perform tests of General Relativity in the time domain, avoiding the spectral leakage problem inherent in the frequency-domain Inspiral-Merger-Ringdown tests.

That the Gaussian process approach models the underlying physical process generating the noise, rather than the noise itself, sets it in contrast to algorithms which model the glitches directly, like BayesWave. For example, in analysing the fast scattering glitch shown in Fig. 4, the Gaussian process approach uses just three parameters: the frequency, amplitude, and quality factor of the simple harmonic oscillator kernel (cf. Foreman-Mackey et al. (2017)). By contrast, BayesWave would need to construct a sine-Gaussian wavelet for each time-separated burst of power. Thus, a GP model can greatly reduce the parameter space size compared to approaches that try to model the glitches directly.

However, the case studies discussed herein amount to only a demonstration of feasibility. There is significant work yet to be done to realise the full potential of the GP approach.

First, we need to extend the scope of the kernel construction. As pointed out in Section 3, we have limited our analysis to the stationary simple harmonic oscillator kernel. But, glitches are a non-stationary process. Therefore, we plan to investigate the use of non-stationary kernels and alternative kernel types, which may better model other glitch classes and enable the analyses of short-duration glitches (relative to the signal duration). In addition, we have applied the pre-whitening approach to circumvent the need to model the stationary coloured Gaussian noise. In future work, we plan to study how to construct a full kernel which can model the coloured noise in addition to transient glitches.

Second, we need to improve the computational and sampling efficiency of our new approach. We can demonstrate the problem by using the analyses presented for the slow-scattering glitch in Case Study A, which required the most complex Gaussian process kernel consisting of three simple harmonic oscillator terms. Averaging over the 10 MCMC chains used in the analysis, the per-likelihood evaluation time of the Gaussian process was $(14 \pm 2)$ ms while it was $(10 \pm 1)$ ms for the Whittle likelihood analysis. This modest additional cost illustrates the remarkable efficiency of the celerite software, without which the Gaussian process would take orders of magnitude more time per evaluation. However, the Whittle likelihood analyses are able to utilise explicitly marginalised likelihoods.
and Walter Del Pozzo, Marta Colleoni, Abhirup Ghosh, and Nathan Johnson-McDaniel for helpful feedback on the time-domain tests of GR. This research has made use of data or software obtained from the Gravitational Wave Open Science Center (gw-openscience.org), a service of LIGO Laboratory, the LIGO Scientific Collaboration, the Virgo Collaboration, and KAGRA. LIGO Laboratory and Advanced LIGO are funded by the United States National Science Foundation (NSF) as well as the Science and Technology Facilities Council (STFC) of the United Kingdom, the Max-Planck-Society (MPS), and the State of Niedersachsen/Germany for support of the construction of Advanced LIGO and construction and operation of the GEO600 detector. Additional support for Advanced LIGO was provided by the Australian Research Council. Virgo is funded, through the European Gravitational Observatory (EGO), by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale di Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by institutions from Belgium, Germany, Greece, Hungary, Ireland, Japan, Monaco, Poland, Portugal, Spain. The construction and operation of KAGRA are funded by Ministry of Education, Culture, Sports, Science and Technology (MEXT), and Japan Society for the Promotion of Science (JSPS), National Research Foundation (NRF) and Ministry of Science and ICT (MSIT) in Korea, Academia Sinica (AS) and the Ministry of Science and Technology (MoST) in Taiwan.

This work makes use of the celerite (Virtanen et al. 2020), matplotlib (Hunter 2007), numpy (Macleod et al. 2021), and PyCBC (Nitz et al. 2017) software for data analysis and visualisation.

REFERENCES

Aasi J., et al., 2015, Class. Quant. Grav., 32, 074001
Abbott B. P., et al., 2016a, Phys. Rev. Lett., 116, 061102
Abbott B. P., et al., 2016b, Phys. Rev. Lett., 116, 221101
Abbott B. P., et al., 2017, Phys. Rev. Lett., 119, 161101
Abbott R., et al., 2021, SoftwareX, 13, 100658
Acernese F., et al., 2015, Class. Quant. Grav., 32, 024001
Ajith P., Bose S., 2009, Phys. Rev., D79, 084032
Akutsu T., et al., 2021, PTEP, 2021, 05A101
Allen B., 2005, Phys. Rev. D, 71, 062001
Allen B., Hua W., Ottewill A., 1999, arXiv e-prints, pp gr-qc/9909083
Ashion G., Talbot C., 2021, MNRAS, 507, 2037
Ashion G., et al., 2019, Astrophys. J. Suppl., 241, 27
Ashion G., Thiele S., Lescoueche Y., McIver J., Nuttall L. K., 2022, Class. Quant. Grav., 39, 175004
Biwer C. M., Capano C. D., De S., Cabero M., Brown D. A., Nitz A. H., Raymond V., 2019, Publ. Astron. Soc. Pac., 131, 024503
Capano C. D., Cabero M., Westerweck J., Abedi J., Kasta S., Nitz A. H., Nielsen A. B., Krishnan B., 2021, arXiv e-prints, p. arXiv:2105.05238
Carullo G., Del Pozzo W., Veitch J., 2019, Phys. Rev. D, 99, 123029
Christensen N., Meyer R., 2022, Rev. Mod. Phys., 94, 025001
Cook S. R., Gelman A., Rubin D. B., 2006, Journal of Computational and Graphical Statistics, 15, 675
Cornish N. J., Littenberg T. B., 2015, Class. Quant. Grav., 32, 135012
Cutler C., Flanagan E. E., 1994, Phys.Rev., D50, 2658
D’Emilio V., Green R., Raymond V., 2021, Mon. Not. Roy. Astron. Soc., 508, 2090
Davis D., Walker M., 2022, Galaxies, 10, 12
Davis D., Massinger T. J., Lundgren A. P., Driggers J. C., Urban A. L., Nitz A. H., 2019, Class. Quant. Grav., 36, 055011
Davis D., et al., 2021, Class. Quant. Grav., 38, 135014
Davis D., Littenberg T. B., Romero-Shaw I. M., Millhouse M., McIver J., Di Renzo F., Ashton G., 2022, arXiv e-prints, p. arXiv:2207.03429
Doctor Z., Farr B., Holz D. E., Pürrer M., 2017, Phys. Rev. D, 96, 123011

5 DATA AVAILABILITY

The data used in the case studies of this work is available through the Gravitational Wave Open Science Center (Abbott et al. 2021).

6 ACKNOWLEDGEMENTS

We would like to sincerely thank Dan Foreman-Mackay for help with the use of the celerite Gaussian process python package. Colm Talbot for the suggestion of the Order Statistics approach to resolving the label-switching degeneracy. Walter Del Pozzo, Christopher Berry, and Andrew Lundgren for useful feedback on the manuscript.
