Transient glitch mitigation in Advanced LIGO data

J. D. Merritt,[1] Ben Farr,[1] Rachel Hur,[1] Bruce Edelman,[1] and Zoheyr Doctor[1,2]
[1]*Institute for Fundamental Science, Department of Physics, University of Oregon, Eugene, Oregon 97403, USA*
[2]*Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA), Department of Physics and Astronomy, Northwestern University, Evanston, Illinois 60201, USA*

(Received 26 August 2021; accepted 11 October 2021; published 15 November 2021)

"Glitches"—transient noise artifacts in the data collected by gravitational wave interferometers like the Laser Interferometer Gravitational-Wave Observatory (LIGO) and Virgo—are an ever-present obstacle for the search and characterization of gravitational wave signals. With some having morphology similar to high-mass, high-mass ratio, and extreme-spin binary black hole events, they limit sensitivity to such sources. They can also act as a contaminant for all sources, requiring targeted mitigation before astrophysical inferences can be made. We propose a data-driven, parametric model for frequently encountered glitch types using probabilistic principal component analysis. As a noise analog of parametrized gravitational wave signal models, it can be easily incorporated into existing search and detector characterization techniques. We have implemented our approach with the open-source *glitschen* package. Using LIGO's currently most problematic glitch types, the "blip" and "tomte," we demonstrate that parametric models of modest dimension can be constructed and used for effective mitigation in both frequentist and Bayesian analyses.

DOI: 10.1103/PhysRevD.104.102004

I. INTRODUCTION

Detecting gravitational waves (GWs) is an immense challenge, requiring the construction and monitoring of the most sensitive interferometers ever built [1]. The strain signal from a loud binary black hole (BBH) inspiral typically perturbs the detectors' arm lengths to one part in 10^{21}. Managing the noise background is an overwhelming portion of that challenge: an earthquake in another hemisphere, a passing vehicle, a cosmic ray hit, a thirsty raven [2], or scattered light from a blinking light-emitting diode can all bring the data well short of the level necessary for detection [3]. In spite of a myriad of obstacles, the LIGO-Virgo Collaboration has detected 58 confident compact binary coalescence (CBC) events as of the end of the first half of the third observing run (O3a) [4,5]. Into O4 these observatories may see confident CBC signals upwards of once a day [6]. Upgrades to the detectors will improve sensitivity and the addition of the Kamioka Gravitational Wave Detector (KAGRA) to the LIGO-Virgo-KAGRA network (LVK) will improve astrophysical parameter estimation (PE) and sky localization. The LVK still expects serious challenges overcoming the noise background, carefully examining more near-threshold triggers, and keeping all the pipelines going with the rapid acquisition of a larger volume of data.

In Advanced LIGO data there are some transient noise sources for which no physical cause has been identified [7]. These noise sources have the potential to impact astrophysical searches significantly [8]. In particular, high-mass and high-mass-ratio BBH searches are affected, in which the astrophysical hypothesis predicts a short-duration signal sweeping up into the sensitive frequency bands of the detectors near merger. Blip glitches and the lower-frequency, longer-duration tomte glitches are glitch types that are capable of masquerading as these high-mass CBCs. These occur on the order of $1/\mathrm{h}$ [7] but sometimes much more frequently, so the probability of coincidence in multiple detectors is non-negligible. Coincident or nearly coincident glitches can confuse search pipelines that strongly rely on coherence between detectors to determine if a trigger is astrophysical. Worse, the effect on the ranking statistic, established by time-sliding data streams from multiple detectors to establish false-alarm rates (FARs) [9], is affected significantly by the presence of these glitches in the background, effectively down-ranking many events. There is evidence that these glitches grow louder and more prevalent with increasing sensitivity [10]. Blips and tomtes all but eliminate our ability to evaluate high-mass, extreme-mass-ratio, and extreme-spin single-detector triggers [11] from a confident astrophysical perspective because the data are contaminated with $\mathcal{O}(10^4)$ loud glitches.

GravitySpy [12] is a pipeline developed to classify glitch types. It leverages citizen science with an image recognition neural network, specifically trained on q transforms, which display power in time-frequency pixels [13]. Thanks to these efforts, there are now over 10^6 glitches classified,

each with an associated confidence metric and SNR [12]. GravitySpy itself can be used to effectively distinguish different types of glitches from each other, but it cannot be used to distinguish signal from glitch, or to subtract glitches from data. For this we seek a parametric, *generative* model for common glitch types. Barring the discovery and mitigation of possible environmental, electronic, or instrumental causes [14–16] for these problematic classes of glitch, distinction between glitchlike astrophysical events and BBH signals that resemble common and problematic glitch types may be our only tractable method for opening up the high-mass and high-mass-ratio region of CBC search parameter space.

With the *glitschen* package, we propose a data-driven, easy to use, and computationally cheap framework for the modeling of short-duration transient glitches. Our model uses an analytical maximum likelihood (ML) estimation approach to fit a probabilistic principal component analysis (PPCA) model to all of the training data, operating under the hypothesis of a transient glitch superimposed on Gaussian noise [17]. While PCAs have previously been used in the context of glitch categorization [18,19], we focus on the construction of glitch-class-specific parametrized models for glitch *mitigation*. Relative to other glitch mitigation techniques [20–24], these targeted parametrized models have minimal flexibility and are in many ways analogous to the parametrized CBC models used to search for and characterize signals, making them straightforward to incorporate in existing LVK analyses. In comparison to current glitch mitigation techniques such as BayesWave (BW) [21], our approach naturally allows for informed priors, allowing us to leverage the extensive glitch population. Our approach can be naturally used in existing analysis libraries such as Bilby, whereas BW's use of reverse-jump sampling means that only a point estimate from BW can be used to remove a glitch during astrophysical parameter inference. While powerful, our methods require large training sets for each glitch type and will likely be unable to model glitches that are extensive in time-frequency, such as scattered light.

II. METHOD

A. Modeling the Advanced LIGO noise background

The noise in the detector is a superposition of many noise sources, and is modeled as a stochastic process, drawing randomly from a stationary background spectrum at each frequency [25]. The detector produces a time series, $n(t)$, which we can represent as a vector, $\boldsymbol{n}$. Transforming to the frequency domain we obtain $\tilde{\boldsymbol{n}}$, with n_i indicating the noise in the ith frequency bin. Assuming Gaussianity, the probability distribution becomes

$$p(\tilde{\boldsymbol{n}}) = \frac{1}{det(2\pi \boldsymbol{C})^{1/2}} \exp\left[-\frac{1}{2}\sum_{ij}(\widetilde{n}_i - \mu)(\widetilde{n}_j - \mu)C_{ij}^{-1}\right], \quad (1)$$

where $\boldsymbol{C}_{ij} = \frac{1}{M-1}(n_i - \mu)(n_j - \mu)$ is the covariance matrix of the observations and μ is the mean of the data [9]. Stationarity means that the noise spectrum is not changing over time, so in the frequency domain the covariance matrix is diagonal: $C_{ij} = \delta_{ij} S_n(f_i)$, giving the power spectral density (PSD), $S_n(f)$ which is equal to the square of the amplitude spectral density (ASD). The noise is typically stationary on the timescales (minutes) relevant for PSD computation, but on the hour timescale may need to be updated [9].

We "whiten" the data by dividing the frequency domain data by an estimate of the ASD, resulting in noise with an equal (unitary) noise in all frequencies. We train and test our model using whitened data.

This treatment is highly effective for "well-behaved" noise sources which remain stationary over the duration of ASD calculation; however, the motivation for building our model is to mitigate transient glitches, which can occur at any time and pose the greatest challenge for searches that look for transient astrophysical events.

B. The transient glitch background

The characteristics of the noise background are well covered in [3,9,10,12]. The morphologies of a typical blip and tomte glitch are explored in Fig. 1. Blip glitches are short, at around 5–10 ms, while tomtes are typically 100 ms long. To properly mitigate these glitches we examine their morphology as they appear to searches, *after* any whitening and postprocessing. Physically, it is possible that the glitches are a very brief dc offset that appears in the strain channel, the result of either a single physical perturbation to some component of the detector or the result of a digital error. We will have to consider the additional morphology of finite impulse response whitening filters as being part of the glitch, since the searches must also contend with these features.

While GravitySpy examines q transforms [13] of glitches, we train on the frequency series of glitches. There is a loss of phase information and direction of amplitude in q transforms, which record only power for each time-frequency pixel. This may be important for future efforts in distinguishing auxiliary witnesses for these glitches, since a preferential directional perturbation to a part of the detector could show up as a bias in amplitude (positive or negative) in the strain channel for a certain detector and glitch type. We have yet to determine if this is a bias introduced in GravitySpy's curation of the highest confidence and loudest glitches, or if this extends to the large number of lower confidence glitches as well, but we see a vast majority of confident, loud L1 O3a tomtes with negative amplitudes. Other detectors and glitch types exhibit a certain "glitch signature" in amplitude bias, sometimes across multiple observing runs.

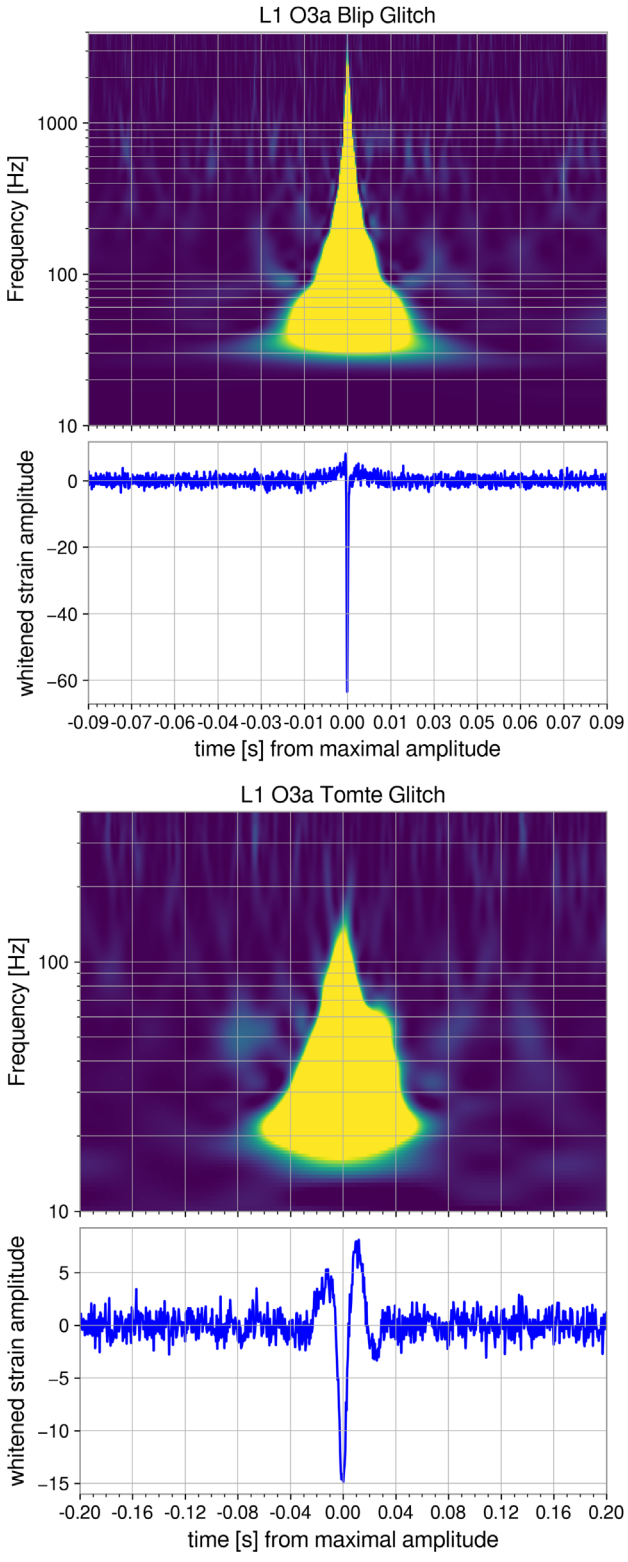

FIG. 1. Typical loud blip (above) and tomte (below) glitches chosen from the test set from Livingston in O3a, to demonstrate glitch morphology. Q scans indicate power in each time-frequency pixel, and the time series (below in blue) shows additional morphology. Note that the timescales and frequency ranges plotted vary. Blips are sometimes shorter than 5 ms, where tomtes can last over 100 ms.

C. The *glitschen* model

In the *glitschen* parametric glitch mitigation model, we employ PPCA. This is a simple and effective way for us to decompose a frequency-domain signal into a set of Gaussian distributed latent variables. It is frequently used as a dimensionality reduction tool, making problems in many areas of data science more tractable. There are many ready-made principal component analysis (PCA) implementations available. We found it most transparent and effective to write our own PPCA implementation, closely following the original PPCA model [17]. This enabled us to find a fast and computationally cheap way to analytically maximize our likelihood. PPCA differs from PCA in that it includes a Gaussian noise term.

We employ an isotropic Gaussian noise model:

$$\mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}), \tag{2}$$

with a d-dimensional observation vector, $\tilde{\boldsymbol{d}}$:

$$\tilde{\boldsymbol{d}} | \boldsymbol{Z}_{train} \sim \mathcal{N}(\boldsymbol{W}\boldsymbol{Z}_{train} + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}). \tag{3}$$

We assume the marginal distribution $\boldsymbol{Z}_{train} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ over q latent variables of the training set, and $\boldsymbol{W}$ has size $d \times q$, containing q training eigenvectors. We recover normal PCA in the limit of $\sigma \to 0$. We can marginalize over the latent variables to obtain a distribution for $\tilde{\boldsymbol{d}}$:

$$\tilde{\boldsymbol{d}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C}), \tag{4}$$

where $\boldsymbol{C} = \boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}$ is the covariance model for the observed data, with dimension $d \times d$. In our case these data are frequency-series data. With N training glitches, our log likelihood for the entire model and all our observed (training) data is then

$$\ln \mathcal{L}_{\text{training}} = -\frac{N}{2}[d \ln(2\pi) + \ln|\boldsymbol{C}| + tr(\boldsymbol{C}^{-1}\boldsymbol{S})], \tag{5}$$

with the sample covariance matrix of the observations, $\boldsymbol{S}$:

$$\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N}(\tilde{\boldsymbol{d}}_n - \boldsymbol{\mu})(\tilde{\boldsymbol{d}}_n - \boldsymbol{\mu})^T. \tag{6}$$

This likelihood is often maximized iteratively, and many packaged implementations of PPCA find W in this way [26]. However we find the global maximum of the likelihood using an analytical method detailed in [17].

Later, we use this likelihood, with an Occam's penalty accounting for the effective degrees of freedom in the model, to find the optimal number of components, q, to use. Performing an eigenvalue decomposition on $\boldsymbol{S}$, the sample covariance matrix of the observations, we obtain the $d \times q$ matrix $\boldsymbol{U}_q$ containing q principal eigenvectors (or "eigenglitches") of $\boldsymbol{S}$, and the $q \times q$ diagonal matrix $\boldsymbol{\Lambda}_q$ with

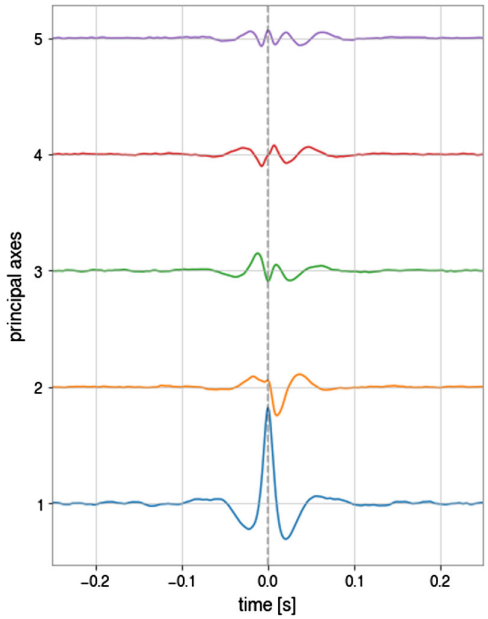

FIG. 2. L1 O3a tomte glitch model eigenvectors. Increasing weight from top to bottom.

corresponding eigenvalues. All eigenvalue decompositions and matrix inversions are conveniently handled by an open-source computer algebra library with NumPy [27]. The likelihood is maximized when

$$\boldsymbol{W} = \boldsymbol{W}_{\mathrm{ML}} = \boldsymbol{U}_q(\boldsymbol{\Lambda}_q - \sigma^2\boldsymbol{I})^{1/2}. \tag{7}$$

Going forward, we can consider $\boldsymbol{W} = \boldsymbol{W}_{\mathrm{ML}}$ to always contain the ML eigenvectors. See an example time-domain representation in Fig. 2.

In order to obtain a projection of a new observation vector, $\tilde{\boldsymbol{d}}_{\mathrm{obs}}$, onto the latent variables we use Bayes' rule to get from $\tilde{\boldsymbol{d}}|\boldsymbol{Z}_{\boldsymbol{train}}$ to

$$\boldsymbol{Z}_{\boldsymbol{train}}|\tilde{\boldsymbol{d}}_{\boldsymbol{obs}} \sim \mathcal{N}(\boldsymbol{M}^{-1}\boldsymbol{W}^T(\tilde{\boldsymbol{d}}_{\boldsymbol{obs}} - \boldsymbol{\mu}), \sigma^2\boldsymbol{M}^{-1}), \tag{8}$$

where $\boldsymbol{M} = \boldsymbol{W}^T\boldsymbol{W} + \sigma^2\boldsymbol{I}$, with dimensions $q \times q$. This allows us to perform reconstructions of suspected glitches using a trained model. We define

$$\boldsymbol{Z}_{\boldsymbol{rec}} \equiv \boldsymbol{M}^{-1}\boldsymbol{W}^T(\tilde{\boldsymbol{d}}_{\boldsymbol{obs}} - \boldsymbol{\mu}) \tag{9}$$

as the set of q latent optimal (i.e., maximum likelihood) reconstruction weights up to an arbitrary rotation matrix. We can obtain a reconstruction with

$$\tilde{\boldsymbol{g}}_{\mathrm{rec}} = \boldsymbol{W}\boldsymbol{Z}_{\boldsymbol{rec}} + \mu. \tag{10}$$

To evaluate the quality of our reconstruction given the data, we employ the standard Gaussian noise likelihood, identical to that used by CBC searches and PE, and specifically Bilby [28]. We define the standard noise weighted inner product of any two frequency-series vectors, $\boldsymbol{a}$ and $\boldsymbol{b}$:

$$(\tilde{\boldsymbol{a}}|\tilde{\boldsymbol{b}}) = 2\int_0^\infty \frac{\tilde{a}(f)\tilde{b}^*(f) + \tilde{a}^*(f)\tilde{b}(f)}{S_n(f)}df. \tag{11}$$

$S_n = \sigma^2$ is the noise PSD, and σ is the ASD [9]. In practice, the noise term can be taken to be 1, because the model is trained on whitened data. When we assume stationary, Gaussian noise that is uncorrelated between detectors, our reconstruction log likelihood becomes

$$\ln \mathcal{L}_{\mathrm{rec}} = -\frac{1}{2}\sum_k \left\{ \frac{|\tilde{\boldsymbol{d}}_{\mathrm{obs},k} - \tilde{\boldsymbol{g}}(\theta)_{\mathrm{rec},k}|^2}{S_k} + \ln(2\pi S_k) \right\}, \tag{12}$$

where k is the frequency bin index, $\tilde{\boldsymbol{g}}(\theta)_{\mathrm{rec},k}$ is the frequency-domain reconstruction with PPCA parameters θ. With this inner product and likelihood we can compare our model's reconstruction of an event, after training on a certain glitch class, with the likelihood of the astrophysical hypothesis. We can select q based on an Occam's penalty, or we can try to replicate the number of effective free parameters in the CBC model to give equal flexibility.

D. Implementation and performance

1. Selection of training data

We curate glitches classified by GravitySpy [12] with high "confidence," where the score ranges from (0,1). Note that confidence is not a normalized probability, but instead reflects the certainty of classification by the convolutional neural network used. We utilize the newest, LVK-internal version of the GravitySpy model, which has the benefit of training on data from all of O3. Publicly available glitch and event data can be obtained from the Gravitational Wave Open Science Center (GWOSC) [29]. This analysis was completed using an older version of the calibrated data: the HOFT_C00 strain data frame within the GDS-CALIB_STRAIN_CLEAN channel. Note that some ($< 1\%$) of the glitches used in training are outside of "science mode" times.

All glitches used first must clear our confidence cutoff (0.95–1, depending on type, detector, and epoch), and are then sorted by SNR. Lower SNR glitches can contaminate the model with more unrelated noise features. As such, we have kept a high SNR threshold for inclusion in training (dependent on type, detector, and epoch), where we use the 1500–2000 loudest glitches. It is more productive to limit the set to "golden" examples curated by GravitySpy, even if the glitch or event in the run segment has low SNR, since we believe quiet and loud glitches (5–50 SNR) exhibit similar morphology, based on our exploration of the data.

2. Preprocessing and training

To train our model, we whiten with an ASD calculated from between 16 and 128 s of data, depending on the glitch type in question. Because we are concerned with the low-frequency content of glitches (in the range of astrophysical searches) all data are downsampled to 2048 Hz, and then for certain glitch types we further bandpass training data to aid in reconstruction efficiency. For Tomte glitches, which have a peak frequency around 50–60 Hz, a 10–128-Hz bandpass to the training data ensures we are not overfitting noise outside the glitch time, but still recover more than 99% of the SNR from more than 99% of training glitches. We find that a 0.5-s training window is always adequate for tomtes, with typical duration 0.1 s. For blips, peak frequencies are typically 500–1000 Hz, so we obtain similar recovered SNR by bandpassing from 10 to 1024 Hz. We find that a 0.1-s window is almost always adequate for blips (allowing one full cycle at 10 Hz). Blips are shorter in duration (almost always shorter than 30 ms). For run segments on test glitches and marginal/glitchlike events we keep data in 10–2048 Hz, retaining higher-frequency noise. All training examples are centered on the peak amplitude time sample. All preprocessing is performed using open-source libraries including NumPy [27] and GWpy [30].

3. Performance

The model is easily run and benchmarked on a laptop with six cores. The training process takes less than 1 s for 2000 glitches. Maximum likelihood reconstruction takes 1 ms − 1 μs depending on how much leeway in center time we allow. Sampling proceeds quickly, giving 10,000 independent samples of the posterior distribution in about 5 min, depending on the glitch.

By weighing the likelihood against an Occam's penalty, we can ensure our model has the appropriate number of dimensions (q) and is not overfitting. We employ a Laplace approximation to the marginal likelihood [31], along with the Bayesian information criterion, described further in the Appendix, to choose the optimal number of eigenvectors for calculating the residuals of the test sets, in the next section. To roughly match the degrees of freedom (per detector) of the CBC model, we employ $q = 5$ in all sampled cases.

III. RESULTS

A. Testing with maximum likelihood reconstruction

We reserve 10% of glitches for testing (the model has never encountered these examples), and to evaluate the performance of our model we examine residuals after maximum likelihood reconstruction and subtraction, as

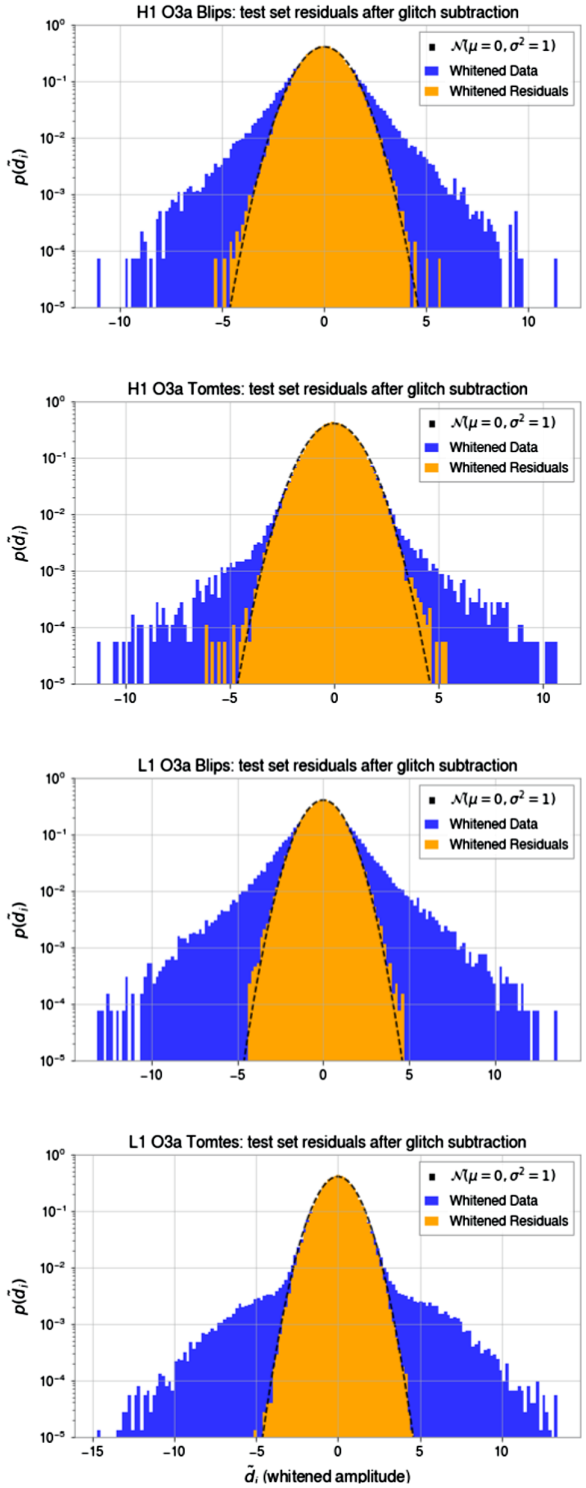

FIG. 3. Frequency-domain residuals after subtraction from the test set (10%) reserved from each glitch type, detector, and epoch. The bins are scaled such that the lowest visible represent single samples from single glitches. Note that extremal samples are louder in Livingston. It has been observed that with greater sensitivity and range transient glitches become louder as well [10].

seen in Fig. 3 histograms. This demonstrates the efficacy of the model in mitigating an entire class of glitches. Test sets shown include 100–200 glitches. We will soon extend this to cleaning entire search backgrounds, and attempt reranking of CBC searches.

We plot residuals after glitch subtraction in the frequency domain. They obey a Gaussian distribution after perfect glitch cleaning under the hypothesis of stationary, uncorrelated noise. Cleaning models are trained with the automatic choice of dimensionality via Laplace approximation (described in the Appendix): (15, 2, 9, 8), for H1 blips (truncated from 23 to 15), H1 tomtes, L1 blips, L1 tomtes, respectively.

The binning in Fig. 3 extends down to single samples from single test glitches, showing that for all classes and detectors our results are consistent with Gaussian noise. The performance is somewhat higher for tomte glitches, mainly due to the greater homogeneity in their morphology compared to blips. Tomtes in Livingston were 10–20 times more prevalent than in Hanford [10]. This has been partly attributed to Livingston operating at greater sensitivity than Hanford during O3a, but may also be due to unknown environmental factors. It is observed that blips and tomtes are louder at higher sensitivity. Higher SNR and greater numbers allow for better modeling, but show the increasing importance of mitigation as sensitivity improves in future observing runs.

B. Sampling

We employ two well-developed Markov chain Monte Carlo (MCMC) toolkits, EMCEE [32], and KOMBINE [33], to perform a full Bayesian posterior estimation of our reconstruction. By allowing the center time of the hypothesized glitch to vary, we sample in $q+1$ dimensions. *A priori*, we assume glitches are equally likely at any time, and thus adopt a uniform prior in center time. The localization of the samples in center time is a good indicator of how glitchlike the morphology of the test signal is. To aid in the efficiency of sampling, we initialize

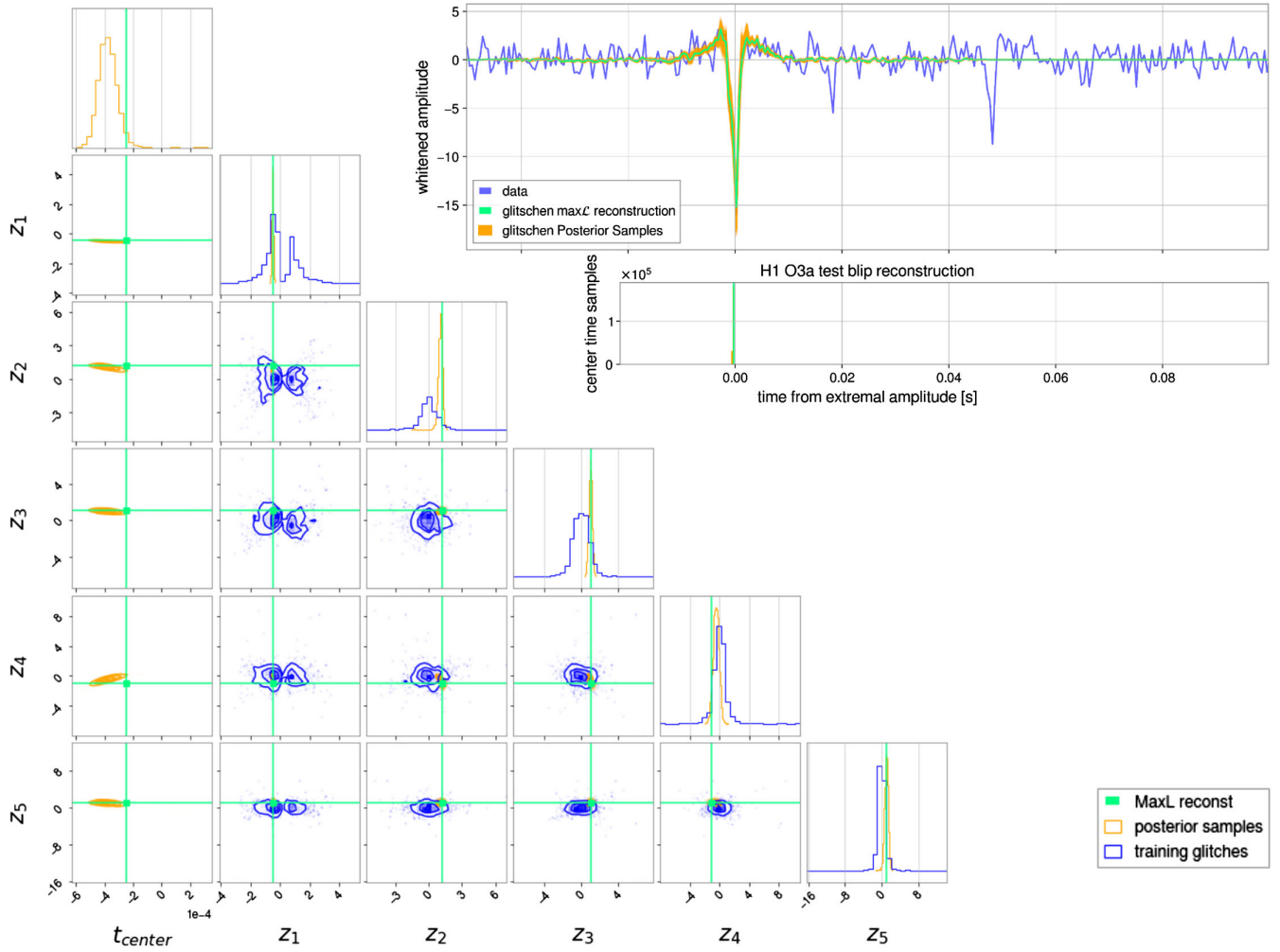

FIG. 4. H1 O3a test blip: full posterior estimation. Note the repeating blips afterward. This example shows the tendency of the sampler to converge on the loudest glitch available. The histogram of center time samples shows high certainty (just below the time-series reconstruction). In the corner plot for the latent space weights (z_1-z_5), we see that this test set glitch is typical of the class.

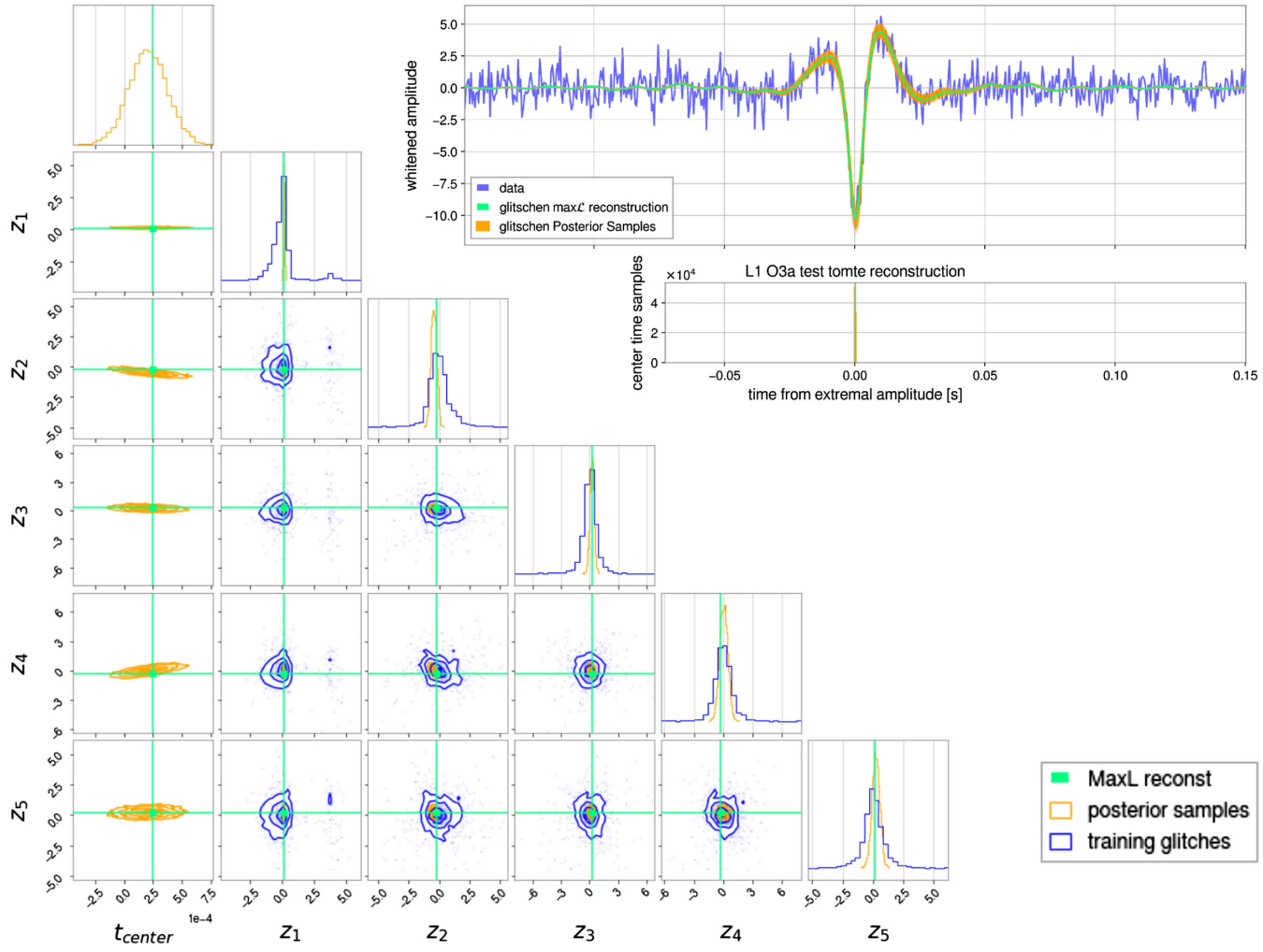

FIG. 5. L1 O3a test tomte: full posterior estimation. This is a very typical tomte glitch, with all walkers converging on the same center time, low uncertainties in the time-series reconstruction, and the posterior distribution aligning well with the distributions on the training set latent weights.

walkers in a Gaussian around the suspected glitch time. In the q PPCA weights, we use a less restrictive wide Gaussian prior, or alternately a highly informed kernel density estimate prior based on the entire training set's maximum likelihood weights. The latter is generally more restrictive and can limit the flexibility of the sampler to fit more general morphologies, which in some cases may be ideal, and in others can be adjusted. For all MCMC sampled

TABLE I. Selecting short-duration, heavy BBH mergers we provide an important test for the model, which should give lower SNRs than the CBC model. Events are in order of detector frame chirp mass ($\mathcal{M}_{\rm det}$, $M_\odot$). For all of these events we see lower SNRs by a factor of 2–3, whereas we expect to recover nearly all of the SNR in confirmed glitches. CBC parameter estimation results from [4].

Event information			Matched-filter SNR					
Event name	$\mathcal{M}_{\rm det}$, $M_\odot$	Duration(s)	CBC H1	Tomte H1	Blip H1	CBC L1	Tomte L1	Blip L1
GW190521	$114.8^{+15.2}_{-17.6}$	0.15	7.87	4.11	3.21	12.38	5.93	4.06
GW190602_175927	$72.9^{+10.8}_{-13.7}$	0.22	6.56	3.60	3.55	11.02	4.43	4.88
GW190706_222641	$75.1^{+11.0}_{-17.5}$	0.15	9.07	4.91	4.22	9.18	3.92	3.93
GW190519_153544	$65.1^{+7.7}_{-10.3}$	0.17	9.50	4.42	4.76	11.85	5.42	4.25
GW190620_030421	$57.5^{+9.0}_{-11.2}$	2.3	(Offline)	⋯	⋯	11.70	3.78	4.63
GW190910_112807	$43.9^{+4.6}_{-3.6}$	1.8	(Offline)	⋯	⋯	13.86	6.29	4.26
GW190521_074359	$39.8^{+2.2}_{-3.0}$	0.24	12.67	5.85	6.30	22.68	8.83	7.24

TABLE II. Running samplers on these events, we obtain the DIC from our distributions of log likelihoods. The DIC favors models with a lower value. The CBC model is highly preferred to the glitch model in all cases, indicating that it passes the signal safety test.

	DIC					
Event name	CBC H1	Tomte H1	Blip H1	CBC L1	Tomte L1	Blip L1
GW190521	−54.6	−5.4	8.7	−130.3	−26.7	9.94
GW190602_175927	−37.3	10.2	8.7	−105.6	−5.6	12.0
GW190706_222641	−71.0	13.2	9.6	−74.0	14.8	8.1
GW190519_153544	−87.7	18.4	−18.6	−145.2	25.1	13.1
GW190620_030421	...	...	...	−133.9	17.5	12.5
GW190910_112807	...	...	...	−190.3	−28.5	17.7
GW190521_074359	−152.8	34.0	36.4	−494.5	−67.1	−45.3

example glitches and CBC comparisons in the paper, we employ $q = 5$.

In Figs. 4 and 5 we demonstrate the results of sampling on a test blip in Hanford, and a test tomte in Livingston. The blip was chosen specifically due to its proximity to further repeating blips. The sampler converges easily on the loudest glitchlike event in the run segment.

C. Signal safety testing

To establish the model's capability of distinguishing glitch from astrophysical signal, we test if it remains flexible enough to fit different glitch morphologies while being (appropriately) unable to reconstruct and subtract an astrophysical signal. We run our model on a selection of high-mass, short-duration BBH signals from GWTC-2 [4],

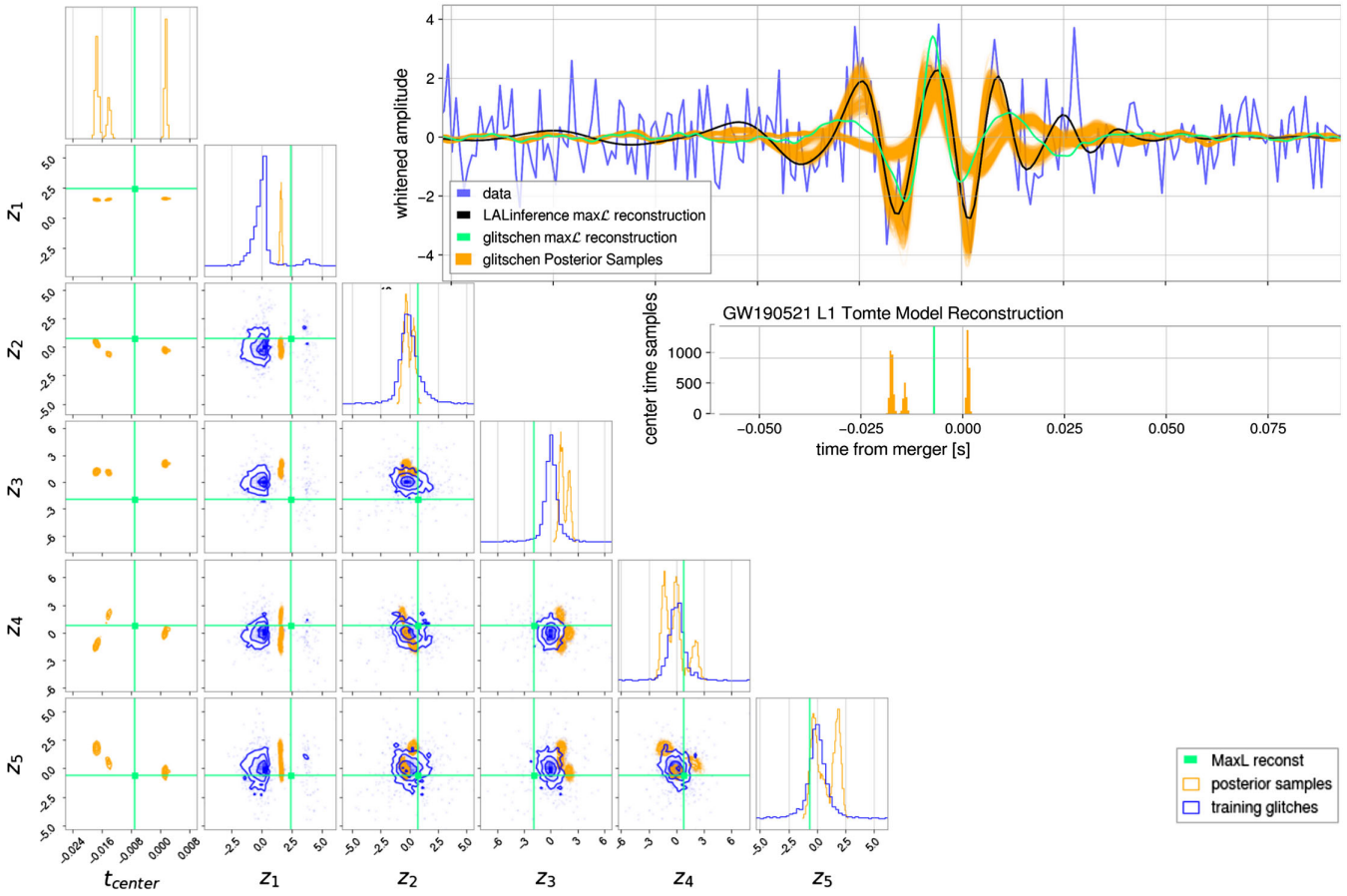

FIG. 6. GW190521 Full Posterior Estimation, L1. The distribution in the center time is multimodal, indicating that the glitch model fails to capture the full morphology of the signal (LALInference maxL in black), no matter where it is placed. The reconstruction features high uncertainty (samples in orange), and the posterior distribution in the latent variables lies outside the training set of glitches. All of this indicates that the model has failed to reconstruct GW190521 as a glitch, as expected.

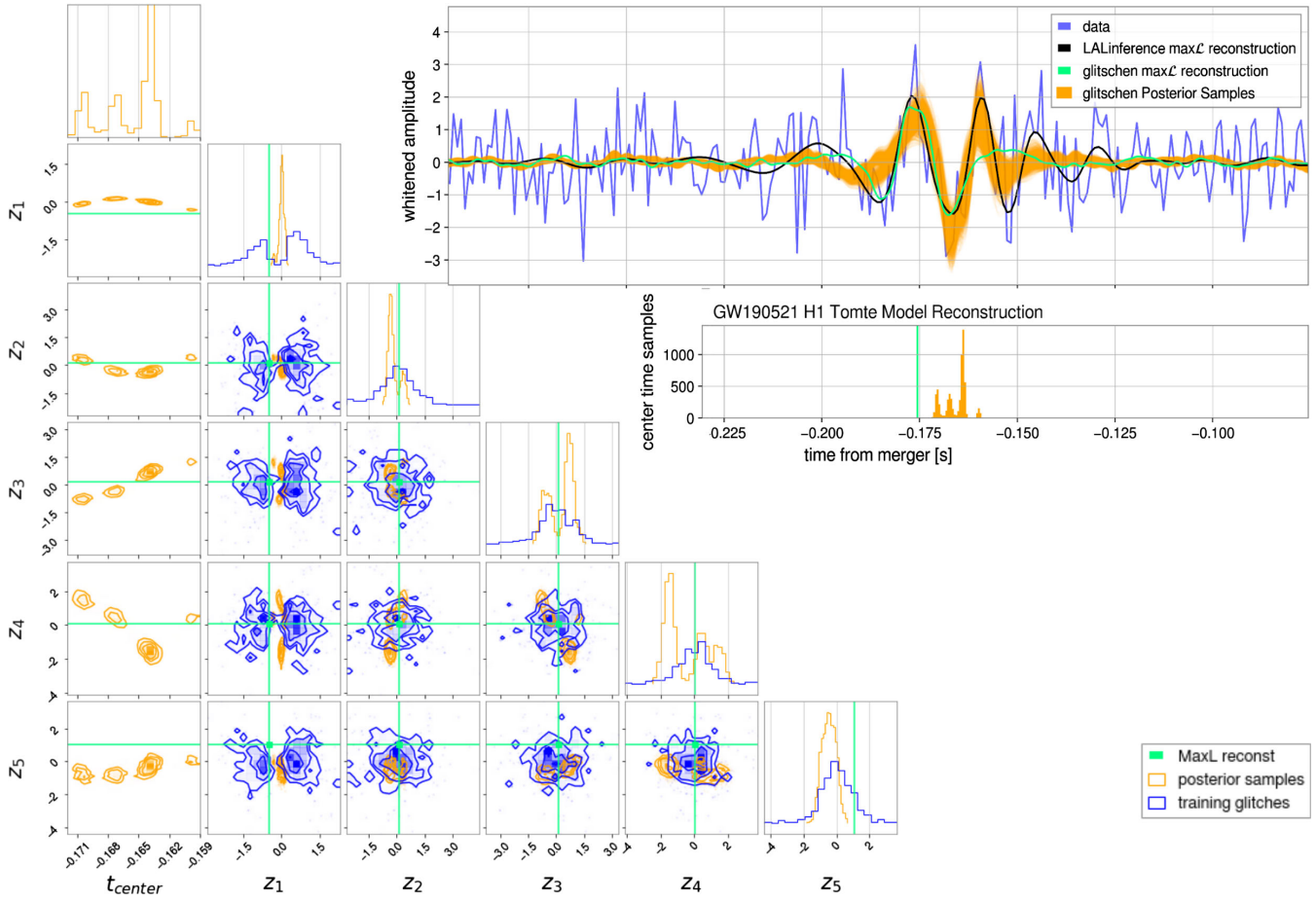

FIG. 7. GW190521 Full Posterior Estimation, H1.

acquiring data from the GWOSC [29]. Specifically, we choose events with high detector frame chirp masses (and by extension, short template durations), and FAR $< 10^{-3}$/yr. Being the confirmed astrophysical signals with morphology closest to short transient glitches such as blips and especially tomtes, these provide a good opportunity to confuse the model. We quote maximum likelihood single-detector SNR values for the CBC and alternately the glitch hypotheses in Table I. Quoted durations are the template duration for the preferred trigger from low latency detection. Two events had H1 offline at the trigger time but were included for their glitchlike morphology. We anticipate that an effective glitch model may be instrumental in vetting single-detector events in the future.

For a more rigorous comparison between the CBC and glitch models we employ a full posterior estimation framework and the Deviance Information Criterion (DIC). This is a variance-based approach. The DIC is given by

$$DIC = D(\bar{\theta}) + \overline{\mathrm{var}(D(\theta))}, \tag{13}$$

where the deviance, D, is $D(\theta) = -2\ln(p(d|\theta))$ with posterior distribution p, data y, and parameters θ. Given the log likelihoods from samples obtained using both the *glitchen* model and a CBC PE run we see that the glitch hypothesis is heavily disfavored for all events tested. These comparisons appear in Table II, where a lower DIC value indicates a better model for the observed data.

D. GW190521: Testing our model's limits with the most massive (and glitchlike) confident O3a event

GW190521 is the highest-mass ($142^{+28}_{-16}\,M_\odot$) and shortest-duration (0.1 s) CBC event for which we have strong evidence [34]. Being a loud triple-detector event, it is confidently of astrophysical origin. But for us, it offers a unique opportunity to test our model, since it exhibits signal morphology which is the most glitchlike of all high-significance astrophysical events. It spent only the last 4 cycles of its inspiral in the sensitive band of the detector, peaking at 60 Hz. Tomte glitches look very similar.

Critically, any model for tomtes, at bare minimum, must not be confused by such an event. Because the aim of improved glitch mitigation is to open up this high-mass region of parameter space, this is precisely the kind of test we need to pass. Here we demonstrate our full posterior estimation framework on GW190521, and by extension, our ability to distinguish glitchlike astrophysical events

from glitches by comparing our glitch hypothesis results with the astrophysical hypothesis results.

In both L1 and H1 (Figs. 6 and 7, respectively), we see that the glitch model (MaxL glitch reconstruction in green) is unable to fully capture the signal morphology, with the MaxL CBC reconstruction in black, no matter where it is placed. It remains multimodal in center time, and an outlier in most of the training set weights, indicating that this is a poor fit to the data, as we expect. High uncertainty is seen in the broadness of the posterior reconstructions, in orange on the time-series plots. See Tables I and II for a more quantitative comparison of the glitch and CBC hypothesis, for this and other events in O3a.

IV. CONCLUSION AND FUTURE WORK

We have introduced a PPCA-based approach to modeling transient noise in gravitational wave detector data, implemented in the open-sourced *glitschen* package, publicly available here: [35]. We welcome collaborative development, testing, and feedback.

For both blip and tomte glitches—some of the most impactful for BBH searches in O3—we have demonstrated the effectiveness of the model for glitch subtraction, as well as for Bayesian model comparisons with astrophysical signal models.

In future work we will explore the use of clustering algorithms in PPCA space for glitch classification and subclassification. We will test the effectiveness of the model in reducing the background for compact binary searches. We will also integrate our model into the Bilby [28] parameter estimation code, where composite signal and noise models will allow us to marginalize over glitch morphology when glitches are coincident with astrophysical signals.

We tested our model on high-mass events from O3a, but in the future we will extend this testing to simulations in the high-mass *and high-mass-ratio* region of parameter space, where discoveries are still to be made and distinguishing astrophysical events from noise is even more difficult.

Because burst searches also trigger on glitches, we plan to test our model in this regime. Searches for cosmic string cusps, supernova templates, and all agnostically unmodeled sources could radically change the field, but only if we can work on the serious blind spots in our searches. We have already began an injection campaign with cosmic string templates in the parameter space contaminated by blip glitches to determine our ability to differentiate signal from glitch in this context. We plan to extend the use of our model beyond Blips and Tomtes, but because these are the most impactful for BBH searches, they remain the first and most important testing ground.

With more accurate models of glitches, we can improve the detectability and significance of gravitational wave events of all kinds.

ACKNOWLEDGMENTS

Here we thank the GravitySpy team, the detector characterization working group, and everyone who made this effort possible. We are grateful to the CBC, Detchar, and Parameter Estimation groups for their hard work building powerful software tools and tutorials and organizing workshops that help make LIGO data analysis work more accessible. The authors are grateful for computational resources provided by the LIGO Lab (CIT, LHO, LLO) and supported by National Science Foundation Grants No. PHY-0757058 and No. PHY-0823459. This material is based upon work supported in part by the National Science Foundation under Grant No. PHY-2110636. We are grateful for the following open-source software tools: NumPy [27], SciPy [36], GWpy [30], Matplotlib [37], and Bilby [28].

APPENDIX: OPTIMAL CHOICE OF THE MODEL DIMENSIONALITY

To avoid over- or underfitting we can use various metrics to find the optimal number of PPCA eigenvectors to employ for each glitch type, detector, and observing run. We tried a crude method: fraction of recovered SNR in test set glitches. If we recover .99 of the known glitch SNR then any gains added with additional dimensions are giving diminishing returns. However this cutoff point is somewhat arbitrary. Instead, balancing an Occam's penalty against the model's training set likelihood is a much more rigorous approach. We employed several methods, including the

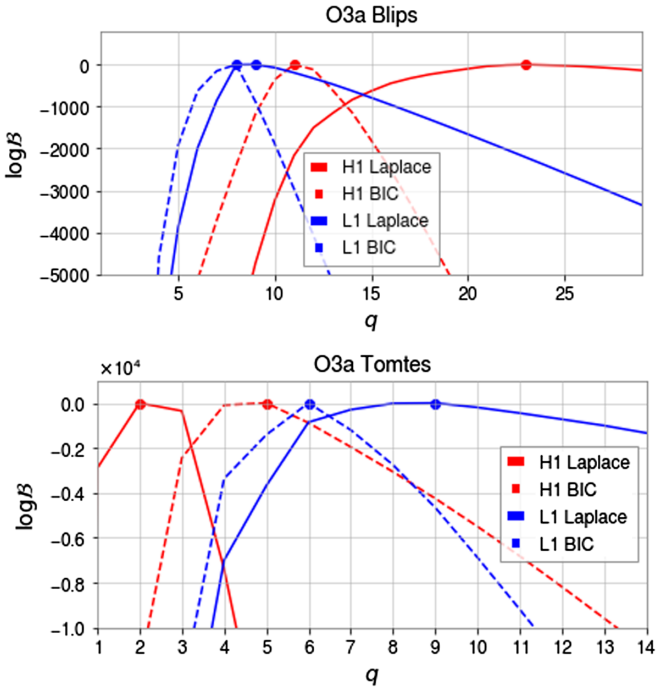

FIG. 8. The relative Bayes factors as a function of dimensionality, q, for each detector and glitch type in the analysis. The peaks of these curves allow for an automatic choice of dimensionality that avoids overfitting.

Akaike information criterion, the Bayesian information criterion, and the Laplace approximation to the marginal model log likelihood, following the method in [31]. By maximizing these metrics we can use the optimal level of model complexity. To arrive at the Laplace approximation, we apply an uninformative conjugate prior on the model parameters and marginalize over everything but q, the PPCA dimensionality. The marginal log-likelihood values are estimates of the model evidence and the ratio of these for different q can be taken as Bayes factors, so far as the Laplace approximation is accurate, which we show in Fig. 8.

[1] B. Abbott, R. Abbott, T. Abbott, M. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. Adhikari *et al.*, Observation of Gravitational Waves from a Binary Black Hole Merger, Phys. Rev. Lett. **116**, 131103 (2016).

[2] R. Schofield, P. Covas, A. Effler, and R. Savage, Why the gw channel detects thirsty black ravens along with colliding black holes (2017), https://alog.ligo-wa.caltech.edu/aLOG/index.php?callRep=37630.

[3] D. Davis *et al.*, LIGO detector characterization in the second and third observing runs, Classical Quantum Gravity **38**, 135014 (2021).

[4] 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 (2021).

[5] R. Abbott *et al.* (The LIGO Scientific and the Virgo Collaboration), GWTC-2.1: Deep Extended Catalog of Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run, arXiv:2108.01045.

[6] B. P. Abbott, R. Abbott, T. D. Abbott, S. Abraham, F. Acernese, K. Ackley, C. Adams, V. B. Adya, C. Affeldt *et al.*, Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA, Living Rev. Relativity **23**, 3 (2020).

[7] M. Cabero, A. Lundgren, A. H. Nitz, T. Dent, D. Barker, E. Goetz, J. S. Kissel, L. K. Nuttall, P. Schale, R. Schofield *et al.*, Blip glitches in Advanced LIGO data, Classical Quantum Gravity **36**, 155010 (2019).

[8] C. Müller and M. Anabel, Gravitational-wave astronomy with compact binary coalescences: From blip glitches to the black hole area increase law, Doctoral thesis, Institutionelles Repositorium der Leibniz Universität Hannover, Hannover, 2018.

[9] B. P. Abbott, R. Abbott, T. D. Abbott, S. Abraham, F. Acernese, K. Ackley, C. Adams, V. B. Adya, C. Affeldt, M. Agathos *et al.*, A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals, Classical Quantum Gravity **37**, 055002 (2020).

[10] S. Soni *et al.*, Discovering features in gravitational-wave data through detector characterization, citizen science and machine learning, Classical Quantum Gravity **38**, 195016 (2021).

[11] D. Davis, L. V. White, and P. R. Saulson, Utilizing aLIGO glitch classifications to validate gravitational-wave candidates, Classical Quantum Gravity **37**, 145001 (2020).

[12] M. Zevin, S. Coughlin, S. Bahaadini, E. Besler, N. Rohani, S. Allen, M. Cabero, K. Crowston, A. K. Katsaggelos, S. L. Larson *et al.*, Gravity Spy: Integrating advanced LIGO detector characterization, machine learning, and citizen science, Classical Quantum Gravity **34**, 064003 (2017).

[13] S. Chatterji, L. Blackburn, G. Martin, and E. Katsavounidis, Multiresolution techniques for the detection of gravitational-wave bursts, Classical Quantum Gravity **21**, S1809 (2004).

[14] V. Braginsky, O. Ryazhskaya, and S. Vyatchanin, Notes about noise in gravitational wave antennas created by cosmic rays, Phys. Lett. A **350**, 1 (2006).

[15] K. Yamamoto, H. Hayakawa, A. Okada, T. Uchiyama, S. Miyoki, M. Ohashi, K. Kuroda, N. Kanda, D. Tatsumi, and Y. Tsunesada, Effect of energy deposited by cosmic-ray particles on interferometric gravitational wave detectors, Phys. Rev. D **78**, 022004 (2008).

[16] A. Helmling-Cornell, Blip glitches in LIGO hanford, https://dcc.ligo.org/LIGO-G2001007/public.

[17] M. E. Tipping and C. Bishop, Probabilistic principal component analysis, J. R. Stat. Soc. Ser. B **61**, 611 (1999); Available from https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/bishop-ppca-jrss.pdf.

[18] J. Powell, D. Trifirò, E. Cuoco, I. S. Heng, and M. Cavaglià, Classification methods for noise transients in advanced gravitational-wave detectors, Classical Quantum Gravity **32**, 215012 (2015).

[19] J. Powell, A. Torres-Forné, R. Lynch, D. Trifirò, E. Cuoco, M. Cavaglià, I. S. Heng, and J. A. Font, Classification methods for noise transients in advanced gravitational-wave detectors II, Classical Quantum Gravity **34**, 034002 (2017).

[20] S. Mukherjee, R. Obaid, and B. Matkarimov, Classification of glitch waveforms in gravitational wave detector characterization, J. Phys. **243**, 012006 (2010).

[21] N. J. Cornish and T. B. Littenberg, Bayeswave: Bayesian inference for gravitational wave bursts and instrument glitches, Classical Quantum Gravity **32**, 135012 (2015).

[22] A. Torres-Forné, E. Cuoco, J. A. Font, and A. Marquina, Application of dictionary learning to denoise LIGO's blip noise transients, Phys. Rev. D **102**, 023011 (2020).

[23] N. J. Cornish, T. B. Littenberg, B. Bécsy, K. Chatziioannou, J. A. Clark, S. Ghonge, and M. Millhouse, BayesWave analysis pipeline in the era of gravitational wave observations, Phys. Rev. D **103**, 044006 (2021).

[24] N. J. Cornish, Rapid and robust parameter inference for binary mergers, Phys. Rev. D **103**, 104057 (2021).

[25] J. D. E. Creighton and W. G. Anderson, in *Gravitational-Wave Physics and Astronomy* (John Wiley & Sons, Ltd, New York, 2011), pp. I–XIV.

[26] J. V. Dillon, I. Langmore, D. Tran, E. Brevdo, S. Vasudevan, D. Moore, B. Patton, A. Alemi, M. Hoffman, and R. A. Saurous, TensorFlow Distributions, arXiv:1711.10604.

[27] S. van der Walt, S. C. Colbert, and G. Varoquaux, The NumPy Array: A structure for efficient numerical computation, Comput. Sci. Eng. **13**, 22 (2011).

[28] G. Ashton, M. Hübner, P. D. Lasky, C. Talbot, K. Ackley, S. Biscoveanu, Q. Chu, A. Divakarla, P. J. Easter, B. Goncharov *et al.*, BILBY: A user-friendly bayesian inference library for gravitational-wave astronomy, Astrophys. J. Suppl. Ser. **241** (2019).

[29] R. Abbott, T. D. Abbott, S. Abraham, F. Acernese, K. Ackley, C. Adams, R. X. Adhikari, V. B. Adya, C. Affeldt, M. Agathos *et al.*, Open data from the first and second observing runs of Advanced LIGO and Advanced Virgo, SoftwareX **13**, 100658 (2021).

[30] D. Macleod, A. L. Urban, S. Coughlin, T. Massinger, M. Pitkin, rngeorge, paulaltin, J. Areeda, L. Singer, E. Quintero, K. Leinweber, and T. G. Badger, GWpy: A Python package for gravitational-wave astrophysics, SoftwareX **13**, 100657 (2021).

[31] T. P. Minka, Automatic choice of dimensionality for PCA, Technical Report No. 514, MIT Media Laboratory Perceptual Computing Section, 2000.

[32] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, emcee: The MCMC Hammer, Publ. Astron. Soc. Pac. **125**, 306 (2013).

[33] B. Farr and W. M. Farr, Kombine: a kernel-density-based, embarrassingly parallel ensemble sampler, https://github.com/bfarr/kombine.

[34] R. Abbott, T. Abbott, S. Abraham, F. Acernese, K. Ackley, C. Adams, R. Adhikari, V. Adya, C. Affeldt, M. Agathos *et al.*, GW190521: A Binary Black Hole Merger with a Total Mass of 150 solar masses, Phys. Rev. Lett. **125**, 101102 (2020).

[35] Https://git.ligo.org/jonathan.merritt/glitschen.

[36] P. Virtanen *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in Python, Nat. Methods **17**, 261 (2020).

[37] J. D. Hunter, Matplotlib: A 2D graphics environment, Comput. Sci. Eng. **9**, 90 (2007).