The Pennsylvania State University
The Graduate School

TOWARDS EXCLUSIVE LOW-LATENCY DETECTION OF GRAVITATIONAL WAVES AND THE INFERENCE OF NEUTRON STAR EQUATION OF STATE WITH NEXT GENERATION DETECTOR NETWORKS

A Dissertation in
Physics
by
Rachael Huxford

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2024

The dissertation of Rachael Huxford was reviewed and approved by the following:

Bangalore Sathyaprakash
Elsbach Professor of Physics and Professor of Astronomy and Astrophysics
Chair of Committee
Dissertation Co-Advisor

Chad Hanna
Professor of Physics and Astronomy and Astrophysics
Dissertation Co-Advisor

David Radice
Assistant Professor of Physics and Astronomy and Astrophysics

Donghui Jeong
Associate Professor of Astronomy and Astrophysics

Mauricio Terrones
Program Head

Abstract

The joint detection of GW170817 and the subsequent kilanova between the LIGO Collaboration, Swift, Fermi, INTEGRAL, and many other observatories heralded our current era of multi-messenger astronomy. With a non-Gaussian transient causing initial issues with the detection of GW170817, this detection proved not only the necessity of prompt detection of such events, but also the importance of low-latency data quality information. However, high-latency gravitational wave searches continue to uncover further detections in archival data which were missed in the initial real-time search at the cost of millions of computing hours.

This dissertation focuses on the improvements to both the GstLAL detection pipeline, and the statistical data quality pipeline iDQ to support the analysis of gravitational waves exclusively in low-latency. This includes discussion of general improvements to both pipelines ahead of LIGO's fourth observing run, and initial results from this period. We then go further to explore new features of the low-latency GstLAL pipeline which both provide live measures of its sensitivity, and increase the sensitivity of its final results without re-filtering any of the data as high-latency pipelines do. We also discuss a new method applied to iDQ which can trace non-Gaussian transients to their sources in the interferometer, and demonstrate how this can be applied in low-latency follow-up of gravitational wave events. Finally, we demonstrate how the sensitivity of future generation of gravitational wave detectors will support the measurement and constraint of neutron star's radii and equation of state with gravitational waves alone.

Table of Contents

Chapter 3

Chapter 4

Chapter 5

Chapter 6

List of Figures

List of Tables

Acknowledgments

I would like to thank my two advisors: B. Sathyaprakash, and Chad Hanna. Sathya, thank you for taking a chance on me and bringing me to Penn State. I never would have made it through grad school without your constant support and belief in me. Chad, thank you for taking me into your group when I was ready to quit and providing me with a path I had never considered before.

Thank you also to the two amazing humans who suffered alongside me through six years: Becca and Divya. Whenever it was unbearable, y'all somehow always made it fun.

A huge thank you to Patrick Godwin for answering my endless questions, and providing a template of what the future could be. This dissertation would never have been complete without your mentorship.

A thank you to James Overduin and Jay Herzog, without whom I would have never even applied to graduate school.

Thank you to my family for encouraging me to do hard things, and follow my passions – even if it meant 22 years of schooling.

Thank you to Randi Neshteruk who was always there with kind words and a guiding hand. Your presence in the physics department is severely missed.

Thank you to Luis Martinez who got me through my qualification exam and taught me everything I know about linear algebra.

Finally, thank you to James Ruth for all the rest. You didn't complain even once when the universe called repeatedly at 2am. Your endless faith and patience carried me through the lowest lows to here, the mountain's peak.

This material is based upon work supported by the NSF under award numbers 1454389, 1642391, 1700765, 1841480, 1841598, 1934752 , 2011865, 2018299, 2103662, 2110594, 2201445, 2308881, 2310912, 2311355, 2346596, 2006384, and 2012083, and by the Indo-US Centre for the Exploration of Extreme Gravity funded by the Indo-US Science and Technology Forum under award number IUSSTF/JC-029/2016. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the NSF or the Indo-US Centre for the Exploration of Extreme Gravity.

Dedication

To anyone who was told they should not, would not, could not.
And yet did.

Chapter 1
Introduction

1.1 Mutli-Messenger Astronomy with Gravitational Waves

With the first detection of gravitational waves in 2015, gravitational wave detectors entered into a world-wide astronomical effort to study the cosmos across multiple observation types. In this multi-messenger ecosystem, the scientific return on any recorded event can only be maximized by combining information from all available diverse observations. Gravitational waves and neutrinos, for example, can travel unhindered from astrophysical sources like binary neutron star mergers (BNS) and carry information from the regions of such mergers which photons and cosmic rays cannot reach. Meanwhile photons and neutrinos can not only create precise localizations of the source environment, but can study the evolution of sources long after gravitational waves in the currently detectable regime have ceased [16].

Additionally, gravitational waves are the first messenger of any merger and, along with any gamma ray burst (GRB), provide initial localization for further observatories to begin searching for any event. However, gravitational waves alone in current detector networks are unable to distinguish between a neutron star and a black hole of the same mass, but the addition of any electromagnetic (EM) counterpart detection would be strong evidence for a neutron star source. Meanwhile, without the localization from gravitational waves, its challenging for EM astronomers to locate compact binary sources so soon after their mergers. As sources like BNS can have a quick evolution in the initial hours, and days following merger, prompt location and observation of these sources is crucial.

Such was the case with GW170817 [17], the first BNS detection by gravitational waves and first multi-messenger observation of such a source. GW170817 was detected first in just a single detector by the LIGO and Virgo collaborations, and was quickly confirmed to be coincident with a GRB at the end of the GW inspiral in Fermi [18]. A data quality issue initially caused this detection to made using Hanford and Virgo data alone, but the full localization of the triple detector network containing Hanford, Livingston, and Virgo narrowed localization in the sky to about 31 square degrees, and addition of the Fermi-GBM and Integral localization further confirmed this. While EM astronomers may have independently observed this source without this localization, the reduction to just 31 square degrees allowed astronomers to locate the source within 10 hours of the initial alert from LIGO, Virgo, and Fermi. Astronomers continued to follow-up on this events in the days, weeks, and even years after the initial discovery. It has only been with a combination of all of these sources that the full evolution of a binary neutron star merger was recorded for the first time.

This highlights the importance that low-latency gravitational wave searches play in the wider global multi-messenger network of observatories. High-latency searches can allow for the comparison of archival gravitational-wave data with other searches such as GRB catalogues, but initial EM radiation evolution from relevant sources would have faded by the time any coincidences are realized. It's only prompt detection of gravitational waves such as GW170817 that give EM astronomers the full notice they need to fully localize and follow-up on these exceptional astrophysical events. This drive for low-latency identification is the foundation for the low-latency GstLAL search discussed in Chapters 2, 3, and 4 and a comparison of the low-latency pipeline's sensitivity with the full offline search given in Chapter 5 shows promising results.

Furthermore, GW170817 highlighted the importance of low-latency data quality information and its incorporation into low-latency search pipelines. If a GW170817-like BNS signal came from a source further away, this event could have initially been ignored because of the data quality issue in the Livingston data. The need for reliable low-latency data quality products is then clear and Chapters 6 and 7 are dedicated to this end.

Finally, even with an SNR above 30, current detector sensitivities did not allow for the accurate measurement of the tidal deformability of the stars in GW170817, nor in any signal since. As will be discussed in the following chapter, the tidal deformability plays such a sub-dominant role in gravitational waveforms that it is extremely difficult to measure accurately. However, future gravitational wave detector networks promise much higher sensitivities, and therefore the radii of neutron stars, along with their tidal deformabilities may be measured by these networks. We show in Chapter 8 that this will lead to constraints on the neutron star equation of state.

1.2 Gravitational Wave Formalism

1.2.1 General Relativity with Gravitational Waves

Gravitational waves are perturbations in spacetime caused by accelerating mass. Taking $g_{\mu\nu}$ as a general spacetime metric, then gravitational waves can be described formally as perturbations on a Minkowski spacetime metric:

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \tag{1.1}$$

where $h_{\mu\nu}$ are the small perturbations in the weak field limit such that $|h_{\mu\nu}| << 1$, and $\eta_{\mu\nu}$ is the Minkowski spacetime metric given as:

$$\eta_{\mu\nu} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1.2}$$

Then to simplify what follows, we impose the Lorenz gauge condition which is equivalent to choosing a set of coordinates. This condition simplifies Einstein's equations such that:

$$(\nabla^2 - \frac{1}{c^2}\frac{\delta^2}{\delta t^2})h_{\mu\nu} = 0 \tag{1.3}$$

This is then equivalent to a wave equation which produces the following solutions:

$$h = A\exp(i(2\pi f t - k_i r^i)) \tag{1.4}$$

where A is the amplitude, f is the frequency, k is the wave number, and the wave is propagating in $\hat{k}$. We then further impose constraints on $h_{\mu\nu}$ equivalent to a transverse-traceless (TT) gauge of:

$$\mathrm{Tr}(h) = 0 \tag{1.5}$$

$$\partial_i h_{ij} = 0 \tag{1.6}$$

$$h_{tt} = h_{ti} = 0 \tag{1.7}$$

$$\tag{1.8}$$

where the indices i and j run over the space components. The use of the TT gauge reveals two polarization states, hereby called the plus and cross states, with amplitudes h_+, and $h_\times$, respectively. These two amplitudes are identifiable when considering our Minkowski spacetime metric with a gravitational wave propagating in the $\hat{z}$ direction where:

$$h_{\mu\nu} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & h_+ & h_\times & 0 \\ 0 & h_\times & -h_+ & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{1.9}$$

From this, we see that the plus polarization imposes an equal and opposite effect in the x and y directions, while the cross polarization is equal in both directions.

Consider a ring of test particles with a diameter given by R, as shown in Figure 1.1. A gravitational wave passing through in the $\hat{z}$ direction with a strain, or amplitude, given by $h = |h_{+,\times}|$ will change the shape of the ring. The ring will first stretch along the x direction by a distance R(1+h), and shrink in the y by R(1-h), thereby forming a flat disk shape. Then, it will go back to a ring before again stretching, but this time in the y and shrinking in the x. Meanwhile, the cross polarization does the same but rotated by 45 degrees, thereby stretching the x and y directions equally making a more diagonal shape. This change in the relative distance of each direction can then be considered as $h = \frac{\Delta R}{R}$ and it is this which is measured by gravitational wave detectors as described further in section 1.3.

1.2.2 Gravitational Waves from Compact Binary Systems

The current ground-based generation of gravitational wave detectors have thus far only detected gravitational waves from compact binary coalescences such as the merger of two black holes

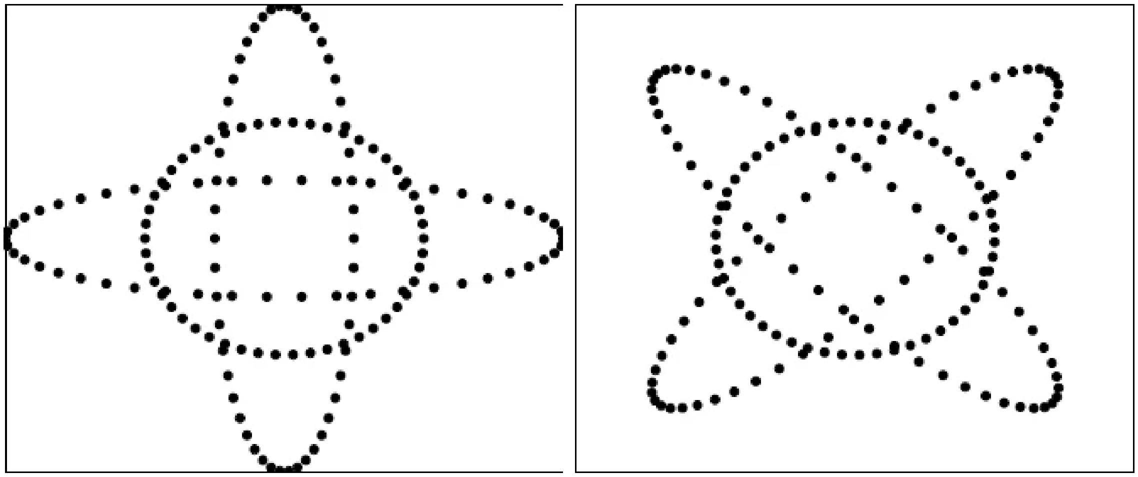


Figure 1.1. The effect of two polarizations of a gravitational wave traveling in the z-direction on a ring of test particles. Shown in (a) the plus polarized wave, and in (b) the cross polarized wave. The rings gets deformed forming one oval of a polarization before returning to the ring state, and then deformed into the opposite oval. Reprinted from [1]

in a binary system, so we will discuss the compact binary system as a source of gravitational waves here.

If we consider the gravitational waves from a compact binary source as a muti-pole expansion, the dominant contribution comes from the quadrupole moment of the system, I_{ij}. Then, in the Post-Newtonian (PN) expansion and considering a locally flat spacetime, the strain is given as [1]:

$$h_{ij} = \frac{2G}{Rc^4}\frac{d^2 I_{ij}(m_k, f)}{dt^2} \tag{1.10}$$

where R is the distance to the source, m_k are the masses of the system, and f is the orbital frequency. Then, for a system of identical masses, m, at a distance r from one another, the strain amplitude is further:

$$|h_{+,\times}| = \frac{32\pi^2 G}{Rc^4} m r^2 f^2 \tag{1.11}$$

This shows that the strain has a the strong dependence on the mass, radius, and orbital frequency of the binary system as well as its distance from the observer. Compact binary sources such as binary black hole (BBH) systems then will be more easily detected compared to binary neutron star (BNS) or neutron star black hole (NSBH) systems which are lighter in mass than a BBH system. The first source of gravitational waves ever detected, GW150914 [19], was in fact a heavy binary black hole system and the first binary neutron star system detected, GW170817 [17], merged within just 40Mpc of us.

The most well-measured parameters in binary systems are those that constitute the leading order terms in the PN expansion of the phase, and therefore appear most strongly in the strain amplitude [20]. Notably, its parameters which are a combination of the individual star's parameters that arise earliest in the PN expansion, and its not until later orders that the degeneracy can be broken. At leading order, the chirp mass, $\mathcal{M}$ plays the leading role and is given as:

$$\mathcal{M} = \frac{(m_1 m_2)^{3/5}}{(m_1 + m_2)^{1/5}} \tag{1.12}$$

At first order, the effective mass ratio comes in and is given as:

$$\eta = \frac{m_1 + m_2}{(m_1 + m_2)^{3/5}} \tag{1.13}$$

At 1.5 order, the effective spin, χ_{eff} of the binary appears, and is therefore well measured, but less so than the chirp mass and can be given as:

$$\chi_{\text{eff}} = \frac{m_1\chi_1 + m_2\chi_2}{m_1 + m_2} \tag{1.14}$$

where χ_i is the dimensionless spin aligned with the angular momentum of each component mass which come at a higher order. It isn't until fifth order that the effective tidal deformability, $\tilde{\Lambda}$ of the system appears as:

$$\tilde{\Lambda} = \frac{16}{13(1+q)^5}\left[(1+12q)\,\Lambda_1 + q^4\,(12+q)\,\Lambda_2\right] \tag{1.15}$$

where $q \equiv m_2/m_1 \leq 1$ is the mass ratio, and Λ_i are the individual tidal deformabilities. The tidal deformability of an object is a measure of how easily it it is deformed in the presence of tidal forces. Black holes, for example, have a tidal deformability of zero, while neutron star's tidal deformability lies in the hundreds to thousands depending on the equation of state used to govern their mass and radius. The individual tidal deformabilities then only appear on their own at the sixth order and are given as:

$$\Lambda_i \propto (\frac{Gm_i}{c^2 r_i})^{-5} \tag{1.16}$$

where m_i, r_i are the mass and radii of the compact objects.

Unfortunately, this results in low-mass black holes being nearly indistinguishable from neutron stars at current detector sensitivities since the tidal deformability components come at such a high order. However, future detector networks may have the sensitivity required to not only measure the tidal deformability well, but distinguish between theorized equations of state governing their density and radii as discussed in Chapter 8.

1.3 Gravitational Wave Detectors

1.3.1 Interferometer Design

Ground-based gravitational wave interferometers are large-scale, modified Michaelson interferometers. The main scientific output of these detectors is the strain data, or the differential difference of length traveled by light down their arms. Each detector has a high powered laser which passes through a beam-splitter and then travels down two perpendicular arms up to 4km in length. The laser light is reflected back at the end, and then recombined with the light from the opposing arm. The light down each arm is kept at a phase shift such that when recombined they exactly de-constructively interfere, resulting in a lack of light on the photodiode monitoring their recombination. As a gravitational wave passes through the detector, it stretches and squeezes the length of these arms thereby resulting in a measurable change in the phase shift and interference of the two laser beams.

We can see this following Creighton and Anderson [21] with support from Saulson [22]. We start with the input laser electric field which follows:

$$E_{in} = E_0 e^{i(2\pi f t - kx)} \tag{1.17}$$

where E_0 is the wave amplitude, k is the wave number, f is the frequency of the light, t is time, and x is distance. The beam splitter is designed to be half reflective and half transmissive, so the magnitude of the power is equal in both arms. This results in reflection and transmission coefficients such that $r = it = \frac{i}{\sqrt{2}}$ and therefore:

$$E_x = iE_y = \frac{i}{\sqrt{2}} E_{in} \tag{1.18}$$

where E_x and E_y are the electric fields for the x and y arm respectively. Assuming the mirrors at the end of each arm have the reflectivity, r_x and r_y, then the fields recombine at the beam splitter where the x arm is reflected and the y arm transmitted. This results in the symmetric and antisymmetric combinations:

$$E_{anti} = \frac{i}{\sqrt{2}}(r_x E_x \exp(i(2\pi f t - kL_x)) + r_y E_y \exp(i(2\pi f t - kL_y))) \tag{1.19}$$

$$= \frac{i}{2} E_{in}(r_x \exp(i(2\pi f t - kL_x)) + r_y \exp(i(2\pi f t - kL_y)) \tag{1.20}$$

$$E_{symm} = \frac{i}{\sqrt{2}}(r_x E_x \exp(i(2\pi f t - kL_x)) - r_y E_y \exp(i(2\pi f t - kL_y))) \tag{1.21}$$

$$= \frac{i}{2} E_{in}(r_x \exp(i(2\pi f t - kL_x)) - r_y \exp(i(2\pi f t - kL_y))) \tag{1.22}$$

$$\tag{1.23}$$

The anti-symmetric combination is the one which leaves the detector and is measured by the photodiode at the output, while the symmetric one is sent back toward the input light. Additionally, under the assumption that $r_x = r_y = 1$, then the anti-symmetric field:

$$E_{anti} = \frac{i}{2} E_{in}(\exp(2\pi f t - kL_x) + \exp(2\pi f t - kL_y)) \tag{1.24}$$

$$= iE_{in} \exp(i(2\pi f t - k(L_x + L_y))) \cos(k(L_x - L_y)) \tag{1.25}$$

$$\tag{1.26}$$

Therefore the power recorded by the photodiode measuring the output which is proportional to E^2 goes as:

$$P_{out} \propto E_{in}^2 \cos^2(k(L_x - L_y)) \tag{1.27}$$

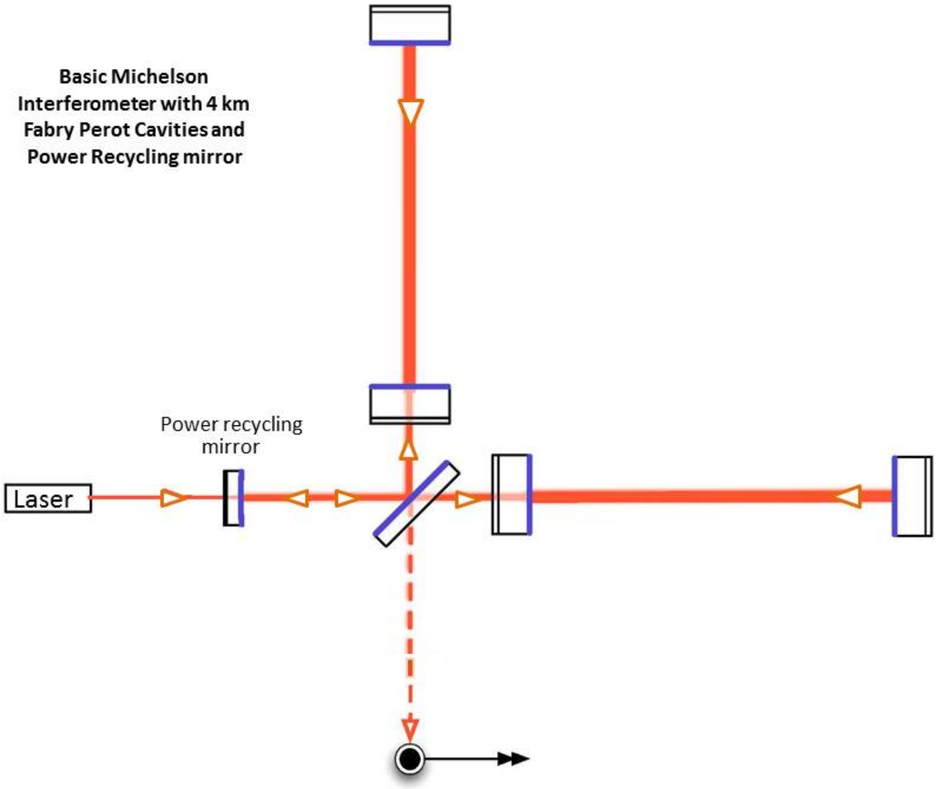

Basic Michelson Interferometer with 4 km Fabry Perot Cavities and Power Recycling mirror
Power recycling mirror
Laser

Figure 1.2. Simplified schematic of the LIGO detectors, from [2], highlighting the Fabry-Perot cavities, and power recycling cavity.

$$\propto \frac{P_{in}}{2}(1 + \cos(2k\Delta L)) \tag{1.28}$$

where $\Delta L = L_x - L_y$ is the difference in the distance traveled by the beams down the two arms.

Returning to the plus polarization of a gravitational wave traveling in the $\hat{z}$ direction through the detector, the x direction would be stretched by $L_x(1 + h)$ and the y would be squeezed by $L_y(1 - h)$. Therefore their difference would become $\Delta L = L_x(1 + h) - L_y(1 - h)$. Assuming $L_x = L_y = L$, then $\Delta L = 2Lh$. It's clear then that the readout of the detector depends heavily on both the input power and the length of the two arms.

In order to increase sensitivity, one or both of these could be increased. The LIGO detectors each have 4km arms, but future upgrades plan on up to 40km ones as discussed in detail in section 1.5. Therefore, other methods have been created to artificially simulate a physically longer arm length without actually requiring the additional space. This is achieved by increasing the path traveled by light in the arm itself, and is the motivation for the design of the Fabry-Perot cavities currently in place. These cavities live in the x and y arms of the detectors and reflect light between two mirrors inside, thereby looping it many times over the length of the cavity before allowing it to be transmitted through the length of the arm. A similar reasoning has motivated the power recycling cavities which live in the path of the symmetric transmitted light. These recycling cavities combine the symmetric transmitted light from the arms constructively with the original input light thereby increasing the power of the input light. A simplified diagram showing the Michelson interferometer and these two additional cavities can be seen in Figure 1.2.

1.3.2 Noise Limitations

The sensitivity of gravitational wave detectors is measured via the range, or the maximum distance a typical BNS source could be measured with a signal-to-noise ratio of 8. The average signal-to-noise ratio (SNR) of a signal as measured in a detector with a single-sided power spectral density, $S_n(f)$, is given as [23]:

$$\langle\rho\rangle = \sqrt{4\int_0^\infty \frac{|\tilde{h}|^2}{S_n(f)} df} \tag{1.29}$$

We can then solve for the distance by Fourier transforming Equation 1.10 for a BNS, plugging it in here, setting $\langle\rho\rangle = 8$, and solving for the distance to the source. This horizon distance is by design for optimally oriented and located sources, so the range is given further for uniformly distributed sources by averaging across sky locations. This results in the range being the horizon distance scaled down by a factor of approximately 2.25.

This range value is then heavily impacted by the power spectral density of the detectors, and various sources of noise change the nature of this function. Despite this, the power spectral density has reached fundamental noise floor limits in some parts of the spectrum, and quickly approaches it in others. These limits can be seen plotted in 1.3.

The laser light incident on the mirrors is one of the main sources of noise in the detector. As light is incident on the mirrors, a small change in the power can change exactly the number of photons hitting the mirror. The variance in this number of photons limits the minimum change in power that is measurable and is called photon shot noise. The contribution to the strain in the detector, or the amplitude strain sensitivity, comes from the square root of the contribution to the power spectral density from this noise [21]:

$$\begin{aligned} h_{shot} &= S_h^{1/2} && (1.30) \\ &= \sqrt{\frac{c\hbar\lambda_{laser}}{2\pi P_{in}}}\frac{1}{L} && (1.31) \\ & && (1.32) \end{aligned}$$

where L is the length of the arms, P_{in} is the input power of the laser, and λ_{laser} is the wavelength. Increasing the power of the laser decreases the shot noise contribution to the overall noise floor, and therefore increases the sensitivity of the detector.

However, unlike the shot noise, radiation pressure noise is made worse by an increase in the power of the laser. At low frequency, the high power of the laser creates radiation pressure noise from the force of the light particles incident on the mirrors. Given a mirror of mass M, the force of the light on the mirrors, or the force from the radiation pressure, is related to the mirror position as [21]:

$$F_{rad} = \frac{2P_{light}}{c} \tag{1.33}$$

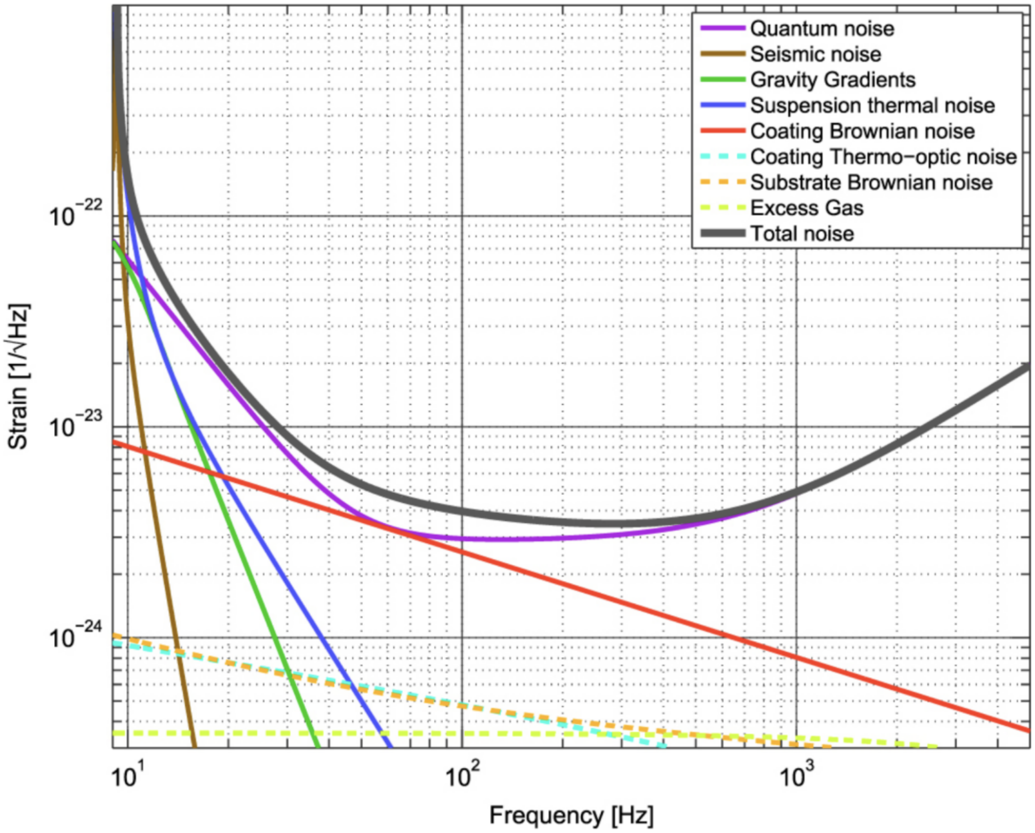


Figure 1.3. Various noise limits from throughout a gravitational wave detector ahead of O3. Reprinted from [3].

$$= \frac{Md^2x}{dt^2} \tag{1.34}$$

$$\tilde{F}_{rad} = -4\pi M f^2 \tilde{x} \tag{1.35}$$

where P_{light} is the power of the light on the mirror. The contribution to $\sqrt{S_n(f)}$, or the strain amplitude, is then given as:

$$h_{rad} = \frac{1}{2\pi f^2 Mc} S_{in}^{1/2}(f) \tag{1.36}$$

$$\propto P_{in}^{1/2} \tag{1.37}$$

where the frequency dependence of this equation makes it dominate at in the low frequency regime. With this, the trade-off between radiation pressure and shot noise becomes clear. The optimal frequency for operation is found by setting the two contributions to the strain equal to one another and solving for the frequency.

The sum quadrature of these two components is labeled as the quantum noise in Figure 1.3 as their smallest values are limited fundamentally by quantum mechanics via the Heisenberg uncertainty principle. These two can be seen then as the dominant noise source in the detector across most of the current noise floor range. However, a few other common noise sources will be mentioned here briefly as next generation detector designs may reach these additional noise floors.

Seismic noise is the noise caused by the seismic activity surrounding the detectors. Sources of this seismic noise can be caused by everything from earthquakes, to the force of the ocean waves and tides on the nearest beaches. However the LIGO Livingston detector in particular is surrounded by forest which is frequently logged. This logging activity does not generally raise the noise floor, but instead causes short duration transient noise during logging periods through the seismic activity.

The coating Brownian noise and suspension thermal noise is caused by the Brownian motion of the molecules on the source of the mirrors and wires used to suspend the optics in the arms. This random motion causes a thermal noise source which peaks at frequencies near the resonance frequency of the mirror suspension systems. Through the right choice of coating material, this effect can therefore be mitigated.

Finally, movements of the earth and air in and around the detector can change their density, and therefore the local Newtonian force on the mirror optics. This is known therefore as gravity gradient noise. By locating detectors underground, the effect from moving air mass can be somewhat mitigated. Otherwise, simply placing detectors in seismically quiet areas is currently the only option for mitigating the effect from earth-based sources for ground-based detectors.

1.4 Data Quality

As described in a previous section, gravitational wave detectors are plagued with a variety of known, constant, noise sources of varying degrees of severity. However, there are additional noise sources which are neither stationary nor of known origin. Short duration non-stationary noise, known as glitches, can have negative affects on gravitational wave search pipelines. Glitches can hinder search pipelines by mimicking real gravitational waves in morphology, obscuring gravitational wave signals, and simply alter the state of the detector noise floor thereby dirtying the background noise distribution used by search pipelines to estimate significance.

Therefore classifying, identifying, and mitigating glitches is of upmost importance to gravitational wave detection, and there are already experts dedicated to this task [24]. Many common types of glitches have been given classifications via their time-frequency morphology in the strain channel. GravitySpy [4,25–28], discussed in more detail in section 6.1.4, is a project built to classify glitches through citizen science efforts, and machine learning. Some of the most common glitch classes used by the GravitySpy team are shown in Figure 1.4. Classifying glitches in this manner can assist in enumerating common appearances, and beginning to identify their sources.

Identifying glitches in the timeseries strain data is also an active area of research. Omicron [29, 30] and SNAX [31], discussed in detail in section 6.1, are both designed to identify non-Gaussian transients in any timeseries data – not just the strain. Both of these are used to locate transients in low and high latency, and have been used for many years by the LIGO and Virgo collaborations.

Once glitches have been located, however, mitigating their affect on the data quality is yet another hurdle for detector characterization experts. There are generally three known ways of handling glitches. First, the source of the glitches could be removed. In order to do so, the source of the glitch must be identified from its morphology in the strain data, or in some other manner. This could be done via an extensive set of hardware injections, where certain subsystems of the detector are artificially stimulated to mimic gravitational waves or glitches. However, performing hardware injections on every subsystem of the detectors is not feasible. Instead, glitches could be traced to certain hardware systems via the monitors on those systems, called auxiliary channels.

There are thousands of auxiliary channels in each detector monitoring the additional degrees of freedom outside of the strain data. These cover the surrounding physical environment monitoring the seismic, acoustic, magnetic, and temperature activity to record, for example, logging activity. They also cover the numerous subsystems within the detector such as those needed to isolate the mirrors from seismic activity, to keep the detector in lock, the temperature on the mirrors, the positions of the mirrors, scattered light off of the mirrors, and more. Sometimes transient noise recorded in these auxiliary channels also appears in the strain data channel as a glitch, and it is these associations which can be used to potentially identify the auxiliary channel source of some glitches, and therefore their origin in the detector. Progress in this direction is detailed in chapter 7.

If the exact waveform of the glitch appearing in the strain data is known, then it could also be subtracted out of the data entirely. The exact form a glitch will take is entirely dependent on the detector state which changes not only between observing runs, but sometimes even between weekly maintenance. There has been progress in this area for glitches which persist across observing runs [32], and modeling of the form these glitches take in the strain data has begun. This would not only allow detector characterization experts to begin attempting subtraction from the strain data, but could also allow detection pipelines to run over large sets of these and determine their performance. However, work in this direction is ongoing and is not yet at the maturity required for active use.

That said, generally gravitational wave detection has been dependent on identifying times of poor data quality, and excluding them from search efforts. As a first order effort, gravitational wave detectors provide state vectors which contain a variety of bits detailing the state of the detector. These state vectors describe when the detector is known to be in a observing and production quality state. Additional information is also contained in the state vector such as details on whether or not hardware injections or other similar maintenance activities were being performed, but the production quality state is never set in the presence of these activities.

In addition to the state vector available in low-latency, sets of time known to contain poor data quality, or vetoes, are created by detector characterization experts in high latency and there are varying levels of these vetoes [33]. CAT-1 vetoes cover times that are truly corrupted by poor data quality, or those times when the detector is generally not considered to be in an observation quality state. CAT-2 vetoes are those which cover times that may contain glitches from well-known sources and/or couplings between the strain data and auxiliary channels. CAT-3 vetoes then cover additional times which may contain glitches from less well established sources and/or couplings. Generally, gravitational wave detection pipelines will use CAT-1 and

CAT-2 vetoes to excise time around known glitches from their analysis set, thereby removing any possibility of this glitch data corrupting their analysis. However, very rarely these times can overlap with true gravitational waves as was the case with GW200129, described in detail in Chapter 3.

1.5 Future Detector Upgrades and Networks

Although described in detail in Chapter 8, the planned detector upgrades will be briefly introduced here. The Advanced LIGO detectors in Hanford, Washington and Livingston, Louisiana are both very close to reaching their design sensitivity goals, and are above-ground detectors with 4km arms. The Advanced Virgo detector located near Pisa, Italy and is also an above-ground detector with 3km arms. These three detectors put together are considered the HLV network, and this network operated during O3, and O4b (HL only during O4a). There are additionally upgrades planned to the LIGO and Virgo detectors, first A+ and then Voyager [34] within the next five to ten years.

LIGO-Anundha in India [35] is a joint effort between the US-based LIGO Lab, and three different institutes in India which recently received funding. The interferometer will be located in India, and will use most of the same hardware as the LIGO Hanford and Livingston detectors. It plans to be a similar above-ground detector with 4km arms, but construction has not yet begun.

KAGRA [36]is the Japanese interferometer, and had plans to join the O4 observing run. Unfortunately, however, an 7.6 magnitude earthquake hit Japan near its location on New Year's Day of 2024. This set back the detector maintenance schedule by about a year, and the current estimation has operations beginning in 2025. The KAGRA detector is unique in the global network as it is set more than 200m underground in an old mining facility beneath Kamioka mountain. It's arms are the same length as Virgo, but they additionally have implemented cryogenic cooling to the mirror systems to reduce the coating and thermal noise discussed previously.

The next generation of ground-based detector networks include the Einstein Telescope (ET) [37–39] in Europe and Cosmic Explorer(CE) [40] in the US. ET is similar in proposal to KAGRA in that it plans to be underground with cryogenics in place for some of the main mirrors. However, it plans to have 10km arms, and possibly have 3 sets nestled inside one another to form a triangle. The CE upgrades will have one 40km detector and one 20km detector, but above-ground and not cryogenic. The main increase in sensitivity for the CE upgrades then comes dominantly from the increase in arm length.

How combinations of these detectors will form advanced network sensitivities is discussed fully in Chapter 8.

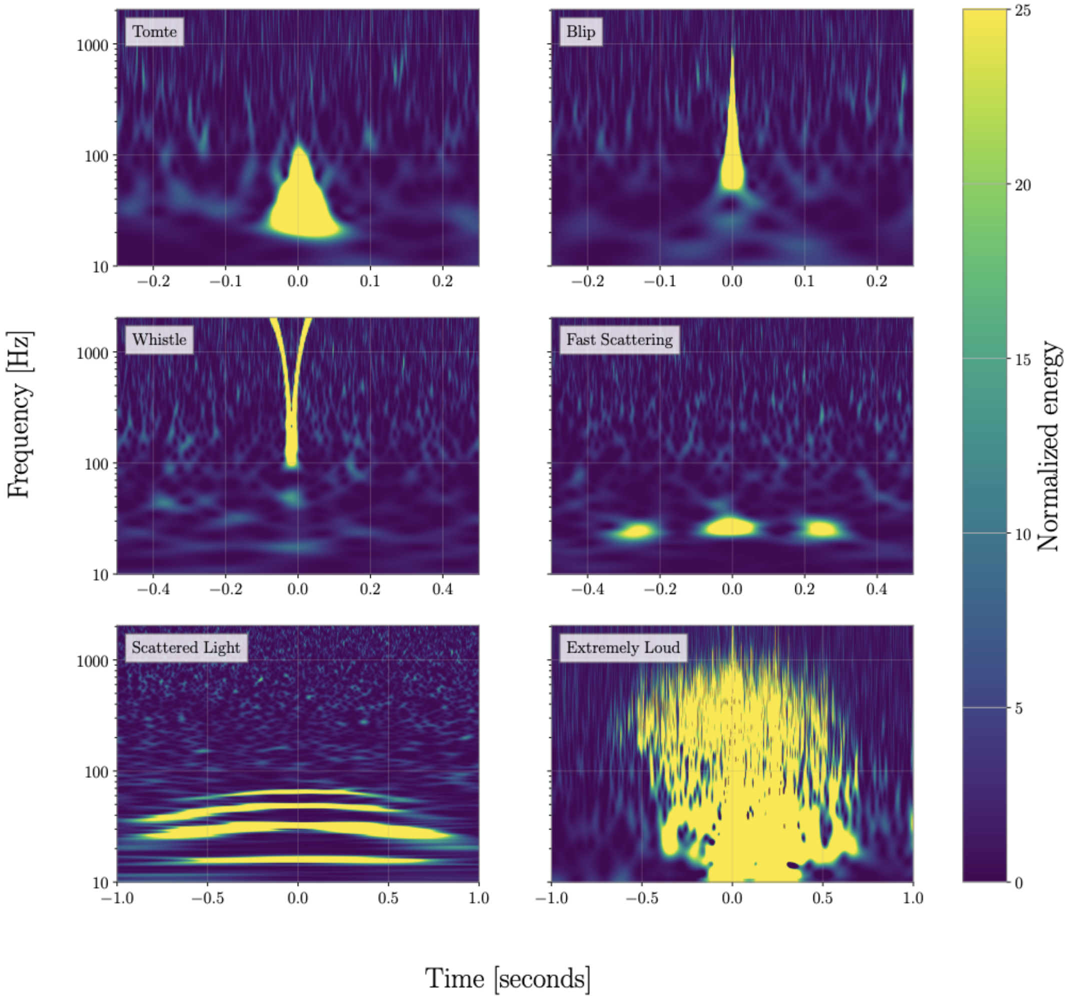


Figure 1.4. Example time–frequency spectrograms for a selection of LIGO glitch classes. The glitch classes here are relatively common and illustrate the range of morphologies different glitch classes can have. The spectrograms in each row are shown with a different time duration. Top left: Tomte is a short-duration glitch with a characteristic triangular morphology. Top right: Blip is another short-duration glitch, but has a tear-drop morphology. Middle left: Whistles have a characteristic V, U or W shape sweeping through higher frequencies (128 Hz). Middle right: Fast Scattering appears as one or more arches, each ∼ 0.2–0.3 s in duration. Bottom left: Scattered Light appears as longer-duration (∼ 2.0–2.5 s) arches, with multiple arches often being stacked on top of each other. Bottom right: Extremely Loud are high-SNR triggers that saturate the spectrogram. Reprinted with permissions from [4]

Chapter 2
Gravitational Wave Search Methods

2.1 Methodology of the GstLAL Gravitational Wave Detection Pipeline

The GstLAL gravitational wave detection pipeline has operated since the LIGO Collaboration began searching for gravitational waves. It relies on matched filtering for the detection of compact binary mergers between the strain data of a detector as:

$$z_i(t) = x_i(t) + iy_i(t) \tag{2.1}$$

where $z_i(t)$ is the signal-to-noise ratio (SNR), $x_i(t)$ is the output of the matched filter with a template of phase ϕ_0, and $y_i(t)$ is the output of the matched filter with a phase-shifted template of $\phi_0 + \pi/4$. The output of the matched filter between the whitened strain data, $\hat{d}(f)$, and a whitened pre-computed waveform (or template), $\hat{h}_i(f)$ with phase ϕ_0 is given as:

$$x_i(t) = 2\int_{-\infty}^{\infty} \hat{h}_i(\tau)\hat{d}(\tau)d\tau \tag{2.2}$$

In the following section, we will discuss the matched-filtering methods implemented by the GstLAL pipeline in detail, and show how the pipeline ranks and calculates significance for possible gravitational wave candidates.

2.1.1 Template Bank Generation & Decomposition

Real gravitational wave signals can come from a variety of sources whose waveforms are defined by their extrinsic and intrinsic parameters. The goal of a matched filtering analysis is to maximize the detection output across all of these parameters. The maximization over the intrinsic parameters, such as masses, spins, etc is done via brute force by match-filtering any data against a large number of simulated waveforms, called templates. Templates are generated at discrete points in the intrinsic parameter space, and the set of these templates which covers the desired search space and which is used in a matched filtering analysis is called a template

O3 Sub-bank Parameter Boundaries					
Label	$m_{1,2}(M_\odot)$	q	χ_1	χ_2	mm
BNS	1-3	0.33-1	low	low	0.99
NSBH	3-150	0.02-1	high	low	0.97
IMBH	3-91	0.1-1	high	high	0.99
low-q BBH	3-392	0.02-1	high	high	0.97
BBH	9-400	0.1-1	high	high	0.99

Table 2.1. The parameter boundaries of the five template sub-banks which combined make the full template bank used during O3. Here, $m_{1,2}$ are the boundaries of the primary masses, q is the mass ratio, $\chi_{1,2}$ are the dimensionless spins, and mm is the allowable mis-match between adjacent templates in the bank. The low and high denotion on $\chi_{1,2}$ represent -0.05 to 0.05 and -0.99 to 0.99 respectively.

bank. How these templates are chosen, organized, and decomposed in a template bank by the GstLAL analysis is described here.

The template bank used through O3 was generated using a stochastic template placing method, described in detail in [41, 42]. The limits of the parameter space for this bank were defined ahead of O3 as follows [43]. The primary masses were limited between 1-400$M_\odot$ generally across the bank. However, the template placement was generated in five distinct parts: BNS, NSBH, IMBH, low-q BBH, and BBH. These five pieces were then combined after the fact to generate the full bank with a total of approximately 1.7 million templates. How each distinct region of the bank was defined can be seen in detail in Table 2.1. In this table, $m_{1,2}$ are the boundaries of the primary masses, q is the mass ratio, $\chi_{1,2}$ are the dimensionless spins, and mm is the allowable mis-match between adjacent templates in the bank. In addition, the high mass region of the bank had extra padding added for the offline analysis to add extra support in the background bins (discussed later in this chapter) in that region. This added an additional approximately 75,000 templates in that region.

The O4 bank was generated using a new software called Manifold [44], which is discussed briefly here. The O4 bank is defined by three parameters: $log(m_1)$, $log(m_2)$, and χ_{eff}. These parameters form a hyper-rectangle in parameter space, and manifold splits this hyper-rectangle down via a binary tree method.

It does this by first defining a metric based on the pre-determined tolerated mismatch between templates in the bank. The local volume in our previously defined three dimensional space can then be calculated from the metric. The number of templates required to cover that space can then be calculated by dividing the total volume by the volume that a template occupies in this space.

Then, Manifold proceeds to define template placements iteratively as follows. First, it defines the initial hyper-rectangle based on the desired coverage in the mass and spin parameter space. Then, it calculates the metric in the center of that rectangle, and the number of templates required to cover it. If the number of templates required is greater than one, it splits the rectangle in half along its longest edge. If the number of templates required is less one, it calculates a template using the parameters at the center of the hyper-rectangle and places it.

It then repeats until the entire space is divided into hyper-rectangles with templates at their centers.

The parameters were defined for the O4 bank [5] as follows . The mass ranges allowed on both masses in the binary was $1 - 200M_{\odot}$ with no breaks. This range was chosen to cover the entire sensitivity range of CBC events expected of the detector network in O4. The mass ratio, m_1/m_2 where m_1 is the heavier component mass, is limited in a range between one and twenty. This is motivated by the fact that during O3, the highest mass ratio event recovered had a mass ratio of 10, so setting a maximum of double that leaves room for new discoveries while reasonably limiting the parameter space of the search. For masses in the range of $1 - 3M_{\odot}$, the dimensionless spin in the direction of the angular orbital momentum, or z-axis, of the binary are restricted to ± 0.05 while those within $3 - 200M_{\odot}$ are limited in a much larger range of ± 0.99. This is because masses in the lower range are expected to correspond to neutron stars. It has been shown that in order for a binary containing a neutron star to merge within a Hubble time (aka those binaries we might expect to detect), it must have a spin less this limiting value of ± 0.05. Black holes, however, can have a spin of ± 1, so the heavier masses are unconstrained. The O4 bank additionally had a minimum template density set in the high-mass BBH region to support a slight over-population compared to the default in that region. This has the same motivation as it did for the O3 template bank – the extra templates is not computationally costly, and supports the background information collection in that region.

The template bank used through O4 with this parameter space contains approximately 2 million templates, or 4 million total waveforms when accounting for the real and imaginary parts. As will be discussed in detail in section 2.1.4, the GstLAL analysis computes the cross-correlation between data and the real and imaginary waveforms individually where a single complex template,$h_i(t)$, is made up of both waveforms. Therefore, the template bank is then doubled by placing both a real waveform, $h_{\times,i}(t)$ and the corresponding $\phi + \pi/4$ phase-shifted imaginary waveform, $h_{+,i}(t)$ at the center of each rectangle such that the complex template:

$$h_i(t) = h_{\times,i}(t) + ih_{+,i}(t) \tag{2.3}$$

Templates in the bank are additionally whitened and normalized such that [23]:

$$1 = 4\int_0^{\infty} \frac{|\tilde{h}_i(f)|^2}{S_n(f)} df \tag{2.4}$$

where $\tilde{h}_i(f)$ is the template waveform, and $S_n(f)$ is the single-sided PSD described in section 2.1.2.

The bank is further broken into two halves in a process called checker-boarding for low-latency analyses. Using Manifold, the smallest hyper-rectangle in any splitting stage is divided into two equal halves, and such provides a natural location for defining the checker-board templates as these two halves will have very similar parameters. In this way, one half of the final splitting of any hyper-rectangle is assigned to one checker-boarded bank, and the other to a second. Thus, we end up with two checker-boarded banks which cover the full input parameter space in almost identical ways, but contain very slightly different templates. Typically, each of these checker-boarded banks is then used to run a low-latency analysis at different computing sites. This prevents downtime at one computing site from disabling the entire low-latency

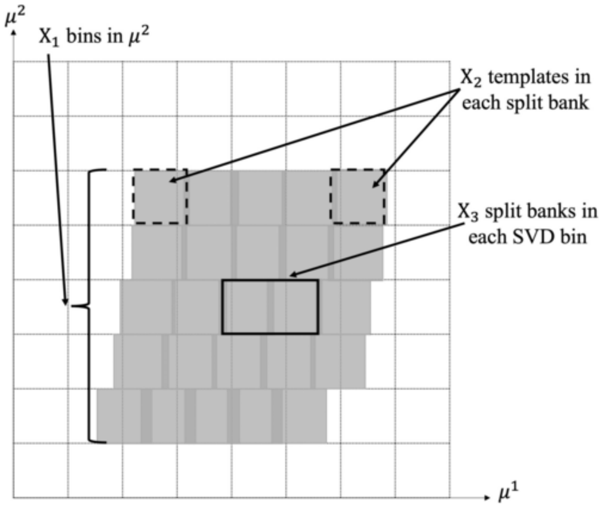

μ^2
X_1 bins in μ^2
X_2 templates in each split bank
X_3 split banks in each SVD bin
μ^1

Figure 2.1. Method of splitting the full template bank into smaller bins, and split banks in the μ_1,μ_2 parameter space where μ_1,μ_2 are linear combinations of the PN coefficients as given by Equation 2.5. In a production analysis, X_1 is typically about 20, X_2 is around 500, and X_3 is typically 2 except for the occasional edge case. Reprinted from [5].

analysis, and the redundancy introduced supports optimal uptime. Its worth noting, however, that only the combined the results from both halves of the full bank constitute the full sensitivity of the analysis.

Still, matched filtering on 2 million waveforms at two different computing sites would be computationally prohibitive, and a two-stage decomposition called the Low-Latency Online Inspiral Detection (LLOID) method was created to reduce this computational load. The first stage of the LLOID algorithm sorts the entire template bank parameter space into bins of templates with similar parameters. This sorting was done in O3 based on two parameters [45]: the effective spin and chirp mass defined as in Equations 1.12, and 1.14. Templates were first grouped by χ_{eff}, and then chirp mass such that about 200 templates were in each split bank.

Ahead of O4, however, this splitting method has been updated to be on two combinations, μ_1, μ_2, of the Post-Newtonian phase coefficients [5], ψ^0, ψ^2, ψ^3 given as:

$$\mu_1 = 0.974\psi^0 + 0.209\psi^2 + 0.0840\psi^3 \tag{2.5}$$

$$\mu_2 = -0.221\psi^0 + 0.823\psi^2 + 0.524\psi^3 \tag{2.6}$$

$$\tag{2.7}$$

Using these sorting parameters instead of chirp mass and effective spin was found to more efficiently group templates by their morphology [5].

As shown in Figure 2.1, templates are first broken into X_1 bins in μ_2, typically 20 in a production analysis. Then, they are grouped into X_3 split banks in μ_1 per X_1 bin where the number of X_3 split banks is defined by assigning an X_2 number of templates per X_3 split bank, typically 500 for a production analysis. Finally, the split banks are grouped together to form approximately 1000 background bins with typically 2 split banks per background bin, although some boundary conditions require the occasional single split bank background bin.

Then, LLOID implements the construction of an orthogonal basis across templates in a given background bin via time-slicing [46]. Each template can be broken into several pieces which are each separated in time, called a time slice. Each time slice in the full template does not overlap whatsoever in time with any of the others, and therefore by construction the full set of time slices for a template are orthogonal in time. Templates with similar parameters, or those in a single background bin, additionally have similar evolution in time-frequency space and are therefore well-suited for similar time slicing boundaries. This allows time slicing boundaries to be set identically for all templates in a single background bin instead of individually per template, thereby further reducing the number of computations.

With a series of time slices in hand, the goal is then to decrease the computational cost of matched filtering any time slice. By default, the sampling rate, and therefore frequency resolution, of the entire template duration is set as f_0, or twice the highest frequency, f_{high} in the template. This sampling rate is defined by the Nyquist rate, or the minimum rate required to avoid aliasing effects at a desired frequency resolution, in this case f_{high}. However, in the quadrupole approximation, the time-frequency relationship is monotonic, $f(t) \propto t^{-3/8}$. Therefore, earlier parts of the waveform have a maximum frequency considerably less than f_{high}, and therefore do not require this high sampling rate. We then down-sample the earlier time slices without loss of resolution. The time slices are defined by their time boundaries $[t_0, t_1), [t_1, t_2), ..., [t_{N-1}, t_N]$, and their sampling rates $f_0, f_1, ..., f_{N-1}$. The sampling rates are twice the highest frequency of any template's time slice in the relevant time range. Any template can easily be reconstructed by lining up their time slices at no loss to the SNR as the SNR timeseries of the full template is just the summation of the SNR timeseries from each timeslice.

A time slice across templates per background bin is then further decomposed into an orthonormal basis by applying a Singular Value Decomposition (SVD) [47]. The decomposition of a time slice for a single template in a bin, $h_i[t]$, is given as:

$$h_i[t] = \sum_{l=0}^{M-1} \nu_{il}\sigma_l \upsilon_l[t] \approx \sum_{l=0}^{L-1} \nu_{il}\sigma_l \upsilon_l[t] \tag{2.8}$$

where $\upsilon_l[t]$ is the matrix of orthonormal basis vectors, σ_l is the vector of singular values which ranks the importance of each basis vector, ν_{il} is the orthonormal matrix of reconstruction coefficients, and the combination of σ_l and ν_{il} gives the reconstruction matrix. The right side of equation 2.8 comes about by truncating the full basis at σ_{L-1} instead of using the full number of bases vectors, M where M is less than the full number of templates in the bin. L can be determined by setting a threshold on the acceptable loss in SNR by truncating the set given by the SVD tolerance (set to 0.99999 during O4):

$$\text{SVD Tolerance} = \left[\sum_{l=0}^{L-1} (\sigma_l)^2\right] \left[\sum_{l=0}^{M-1} (\sigma_l)^2\right]^{-1} \tag{2.9}$$

This decomposition allows the matched filter to be computed L times per time slice instead of a number of times equal to the number of templates in the bin. The combination of sorting, time slicing, and an SVD basis then vastly reduces the number of filters required from the original number of templates at a minimal loss of SNR.

2.1.2 PSD Estimation & Whitening

Matched filtering algorithms require the single-sided noise power spectral density, $S_n(f)$ to whiten the data and templates. Complications immediately arise in the gravitational-wave detection use-case thereby limiting how this value is measured and estimated. As mentioned previously, gravitational wave noise data frequently contains short duration departures from the typical noise background. Gravitational wave signals can also appear in the data at both short and medium duration depending on their source properties. In both cases, we do not want the PSD to be polluted by the presence of the departures from the typical noise background. We follow here the logic of the GstLAL analysis' measurement of the PSD given originally in [23] with additional support as necessary.

Typically, the PSD can expressed as a function of the Fourier transform of the noise $\tilde{n}(f)$ as [23]:

$$\langle \tilde{n}(f)\tilde{n}^*(f) \rangle = \frac{1}{2} S_n(f)\delta(f - f'), f > 0 \tag{2.10}$$

where the half arrives from using the single-sided PSD, and the $\langle ... \rangle$ is the ensemble average. However, Equation 2.10 will be negatively affected by the aforementioned presence of gravitational waves and glitches. Additionally, the PSD of the detector may slowly drift over time-scales shorter than an observing run or lock segment, and we would like to capture this behavior. Instead, a combination of the median and geometric mean is used to mitigate the issues that arise from glitches and gravitational waves, while additionally chunking historical data into reasonably sized pieces for suitable analysis in low-latency as described in full detail in [23].

We begin by chunking historical timeseries data in pieces with N samples which overlap by N/2 + Z samples where N and Z must be even and Z is typically taken to be N/4. A Discrete Fourier Transform (DFT) assumes that the data is infinitely periodic, which is not true for our chunked data as the data at the left boundary is not guaranteed to match the data at the right of our arbitrary boundary choice. If the DFT is applied to the data as is, it will appear to the DFT to have periodic discontinuities mimicking the edges of a square wave, thereby creating false lines, or clicks, in the frequency domain which can themselves mimic the shape of gravitational waves. Therefore, we first apply a Hann window to these N samples, suppressing the discontinuities at the boundaries by forcing the series to zero at each end. When Z=0, the window function $w[k]$ is given as:

$$w[k] = \sin^2(\pi k/N) \tag{2.11}$$

where $k \in [0, ..., N-1]$ is the time sample index.

To appropriately overlap chunks with the Hann window when Z=0 such that they sum to unity everywhere, the middle sample of the first chunk, where the function is one, must overlap with the last sample of a future chunk, where the function is zero. However, N is enforced to be even for the DFT, meaning there is no odd middle sample. The chunk is therefore padded by a single sample, the window is applied, and then the last sample is dropped. The result can be seen in Figure 2.2.

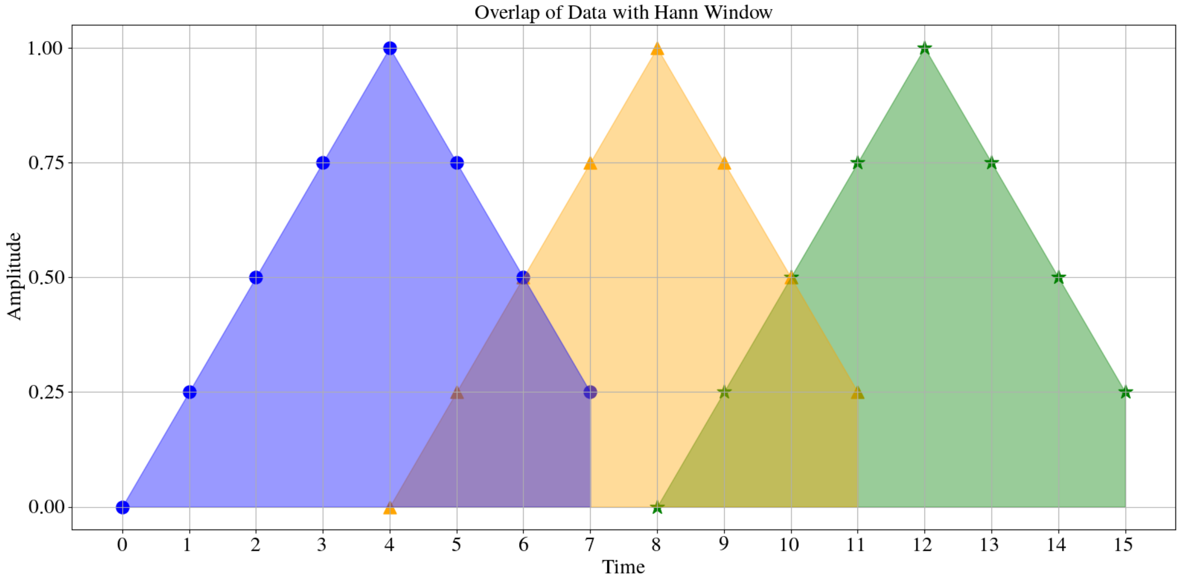


Figure 2.2. Cartoon overlap of the Hann windows used with timeseries data to suppress discontinuities at the boundaries. The last sample of each window is dropped, so an even number remains for the DFT. The sum everywhere is left as unity.

To further mitigate any remaining boundary issues and to increase the frequency resolution, the chunk is then additionally zero padded by a number of zero samples equal to Z such that the window function becomes:

$$w[k] = \begin{cases} 0, & k < Z \\ \sin^2(\frac{\pi(k-Z)}{N-2Z}) & Z \leq k < N - Z \\ 0, & N - Z \leq k < N \end{cases} \tag{2.12}$$

The full application of the window and DFT to a given chunk of timeseries data, $d_j[t]$ can then be given as [23]:

$$\tilde{d}_j[\ell] = \sqrt{\frac{N}{\sum_{k=0}^{N-1} w[k]^2}} \Delta t \sum_{k=0}^{N-1} w[k] d_j[k] \exp(-2\pi i \ell k/N) \tag{2.13}$$

where $\ell \in [0, N/2]$ is the frequency bin, and Δt is the time sample step. Additionally, the window function goes to zero at both zero and N/2, so $\tilde{d}_j[0]$ and $\tilde{d}_j[N/2]$ are set to zero. Finally, the PSD can given as:

$$S_j(f) = 2\Delta f |\tilde{d}_j[\ell]|^2 \tag{2.14}$$

with a frequency resolution $\Delta f = 1/(N\Delta t)$.

Recall, however, the pollution of the PSD by glitches and gravitational wave signals needs to be mitigated via the implementation of the median and geometric mean. We take the median of the last n_{med} chunks then as:

$$S_j^{med}(f) = \text{median}\{2\Delta f |\tilde{d}_n[\ell]|^2\}_{n=j-n_{med}}^{n=j} \tag{2.15}$$

The median functionally ignores outliers, so if an event like a glitch caused issues during a single chunk for example, the median would be unaffected. Any polluting event would have to last for at least half of the chunks covered by the median, or a time period of $\frac{1}{2}n_{avg}(N/2)\Delta t$, to shift the median in any way. With the typical value of $n_{med} = 7$, this corresponds to more than 3 seconds, which is much longer than the average glitch duration, but not necessarily longer than a gravitational wave event.

To account for gravitational-wave event timescales, the geometric mean is introduced. Under the assumption that the detector noise is stationary and Gaussian, the frequency bins in the PSD will be randomly drawn from a Gaussian distribution and are therefore χ^2-distributed. The geometric mean of any n_{set} set of variables is defined as:

$$\langle a_l \rangle = \exp\Big(\frac{1}{n_{set}} \sum_{i=0}^{n_{set}-1} ln(a_l)\Big) \tag{2.16}$$

where a_l is the set, and $i \in [0, ..., n_{set} - 1]$ is the index of the set. Applied to our use case, $n_{set} = N$ for any given chunk. As chunks arrive in time, we update the running geometric mean by weighting the geometric mean of any previous chunks by $\frac{n_{avg}-1}{n_{avg}}$ and the geometric mean of the most recent chunk by $\frac{1}{n_{avg}}$ where n_{avg} is the number of chunks up to this point. In other words:

$$\langle a \rangle = \exp\Big(\frac{n_{avg} - 1}{n_{avg}} \langle a_{0,...,l-1} \rangle + \frac{1}{n_{avg}} \langle a_l \rangle\Big) \tag{2.17}$$

where $\langle a_{0,...,l-1} \rangle$ is the geometric mean of all the chunks except the latest one, $\langle a_l \rangle$ is the geometric mean of the latest chunk, and $l \in [0, ..., n_{avg}]$ is the index of all the chunks considered, and notably not the index of the samples in a single chunk.

The geometric mean of a χ^2-distributed variable is conveniently equal to the median divided by a proportionality constant β. Then, Equation 2.17 becomes [23]:

$$S_j[\ell] = \exp\Big(\frac{n_{avg} - 1}{n_{avg}} \langle S_{j-1}[\ell] \rangle + \frac{1}{n_{avg}} \frac{S_j^{med}[\ell]}{\beta}\Big) \tag{2.18}$$

where $j \in [0, ..., n_{avg}]$ is the chunk index and, again, n_{avg} is the number of chunks tracked by this running average, typically 64. As chunks become old enough to no longer be in the time range of n_{avg} chunks, their data is dropped from the running mean.

Finally, the arithmetic mean is estimated from the geometric mean and used to whiten the templates and data. The arithmetic mean of a χ^2-distributed variable is simply its geometric mean by e^γ where γ is Euler's constant. Then, in this discrete case, a template or stretch of data in the frequency domain, $\tilde{d}_j[\ell]$ can be whitened via:

$$\hat{\tilde{d}}_j[\ell] = \frac{\tilde{d}_j[\ell]}{\sqrt{S_j[\ell] \exp(\gamma)}} \tag{2.19}$$

where the hat denotes whitening, and where $S_n(f) = S_j(f)e^\gamma$ is the representation of the PSD in the continuous case. We can then represent the full whitened data stream in the time domain as [23]:

$$\hat{d}_j[k] = 2\Delta t\Delta f \sqrt{\sum_{n=0}^{N-1} w[k]^2} \sum_{\ell=0}^{N/2} w[k]\hat{\tilde{d}}_j[\ell] \exp(2\pi i k\ell/N) \tag{2.20}$$

where all of the additional terms ahead of the summation are required for normalization. Ahead of the fourth observing run, the values of N, and therefore Z were updated in the low-latency analysis to $N = 4f_s$, and $Z = 1f_s$ where f_s is the sampling rate. These values were lowered from their previous settings in order to decrease the latency incurred by the whitener to mere 2 seconds. Additionally, templates in the template bank were previously whitened once at the start of an observing run, but during the fourth observing run, templates were re-whitened every week based on the PSD calculated from the previous week's data in order to keep them as up-to-date as possible.

2.1.3 Data Conditioning

As introduced in section 1.4, detector data quality is monitored at the coarsest grain by detector state vector channels which can provide simple information about the state of the detector during data collection. Frequently, however, even when in lock and in a science-quality state, the data is not always stationary and Gaussian as we typically assume. While slow, long-duration changes are tracked by the PSD as discussed in section 2.1.2, the PSD was specifically designed to be agnostic to short duration variation from Gaussianity. Instead, the detector data can still be affected by glitches, or short duration noise transients which vary from the usual background noise levels. There have been a variety of methods employed by different detection pipelines to mitigate the effect glitches have on detection confidence, and the implementation in GstLAL is discussed here.

In the GstLAL pipeline, times of poor detector quality are handled in two ways: by removing them entirely from the data ahead of matched filtering, or by mitigating their effect on confidence statistics after the fact. The latter is discussed in section 2.1.6, while the former is described here. Once the data has been whitened, it should have unit variance, but introduction of glitches changes this. When glitches cause departures that are greater than some number of standard deviations, σ_{thresh}, GstLAL excises them via a process called gating. During gating, some window in time around the glitch is completely removed from the analysis, that is, the data is identically set to 0 in the entire region. It's assumed that glitches which cause the most extreme departures, and are therefore most likely to mimic a high-mass CBC gravitational wave signal, are well-localized in time to short durations. During the whitening process, noise such as this can have its power distributed to a wider duration in time than its original locality, however, a result called spectral leakage. It has been shown that the typical whitening filter used by the GstLAL analysis can be approximated as a narrow sinc function in the time domain with about 98% of its magnitude contained within 10ms [48]. Then, the probability of large amounts of spectral leakage in a wide window around the original glitch from the application of the whitener is generally small. Therefore, a quarter second on either side of a signal passing the threshold is removed which is a conservative estimate under the assumption of short duration transients.

Additionally, as its known that heavier-mass CBC systems tend to more closely resemble the strongest glitches, the GstLAL analysis scales the gating threshold as a linear function of chirp mass. This is an effort to reduce the number of real astrophysical signals that might be removed by this gating process, and so the thresholds are set conservatively on a per background bin basis using the follow scale [9]:

$$\sigma_{\text{thresh}} = R_{\text{thresh}} \Delta \mathcal{M}_c \sigma_{\text{min}} \tag{2.21}$$

where $\Delta\mathcal{M}_c$ is the difference in chirp mass between the highest and lowest chirp mass in a single background bin, σ_{min} is a tunable minimum threshold typically set as 15, and R_{thresh} is the tunable scaling factor given as a little less than 2 during O4.

Additional times can also be flagged in high-latency as having poor data quality as described by the CAT-n veto system discussed in 1.4. This is typically done via the analysis of relevant auxiliary channels, or by times logged after the fact by detector operators. These times are collated into a set of detector quality vetoes, and during high latency analysis, the GstLAL pipeline will apply these vetoes to the data. Instead of applying them to the input timeseries, however, they are applied by setting the whitened timeseries to zero in their durations. As whitened data is defined as a series of uncorrelated samples, this can be done without any negative affect to the analysis.

2.1.4 Matched Filtering

The goal of a matched filtering analysis is to maximize the output over the intrinsic and extrinsic parameters of a source. The extrinsic parameters for aligned spin waveforms come out as an amplitude factor for the overall waveform, but the intrinsic parameters require more care even for the simple cases.

The intrinsic parameters are maximized over using the brute force method of maximizing matches over a template bank. The time of coalescence is maximized by taking the maximum matched-filter output from that bank in some small window. The result of a matched-filter is then the SNR given as the cross-correlation between a single whitened template, $\hat{h}_i$, and the whitened data, $\hat{d}$. For GstLAL, it is calculated in the time domain as:

$$x_i(t) = 2 \int_{-\infty}^{\infty} \hat{h}_i^*(\tau)\hat{d}(t+\tau)d\tau \tag{2.22}$$

where the hats denote whitening, and the discrete whitening process is given for a template, $\hat{h}_i$, in equation 2.19, and the discrete whitening process for the full data stream, $\hat{d}$, in equation 2.20.

However, recall that the template bank is further decomposed using the LLOID method such that one time slice of a single template in a given background bin is given by Equation 2.8. Then, the matched-filter output must additionally be calculated per time slice, s, and per SVD basis vector, υ_l, as:

$$U_l^s(t) = 2 \int_{-\infty}^{\infty} \upsilon_l(t)\hat{d}(t+\tau)d\tau \tag{2.23}$$

where υ_l is the basis vector as defined in 2.8, and the hat again denotes whitening. The output across basis vectors for a single time slice can then be reconstructed using the reconstruction matrix following 2.8 with $U_l(t)$ in place of ν_l as [23]:

$$x_i^s(t) = \sum_{l=0}^{L-1} \nu_{il}^s \sigma_l^s U_l^s(t) \tag{2.24}$$

Each time slice is then upsampled back to the original sample rate, and lined up in time order end to end to reconstruct the full template's SNR timeseries as given in 2.22.

As discussed in section 2.1.1, for each set of intrinsic parameters, the GstLAL bank additionally has two waveforms, one for the real and one for the imaginary parts corresponding to the plus and cross correlations. The matched-filter output of the two are then combined as:

$$\rho_j(t) = \sqrt{x_j(t)^2 + y_j(t)^2} \tag{2.25}$$

where $x_i(t)$ is the result of 2.22 for the real part, and $y_i(t)$ the same for the imaginary part, and j is the template index.

This complex SNR timeseries is of a sample rate of 2048Hz, resulting in a data size far too large to store on disk for any reasonable amount of time. Therefore, we additionally maximize over the unknown time and phase of the signal by maximizing the SNR timeseries per template in one second windows. If the SNR in that window crosses some predetermined threshold, typically 4, then it is stored to disk as a trigger. Triggers have their associated template parameters, time of coalescence, SNR, and coalescence phase stored.

To form potential gravitational wave events, triggers are compared across detectors, and those who have at least one pair within a small time window are considered an event. The time window used is based on the light travel time between the two detectors with a bit of additional padding due to noise fluctuations in the detectors, and it is typically taken to be ±5ms. If there is a coincidence, the template from the loudest SNR trigger across all detectors is then paired with the results from the same template in the other detectors. The template which had the loudest SNR across detectors may not have had the loudest SNR in each detector, but we still take that template's results from each. Although this could result in a small loss of SNR, keeping the results with the same origin template parameters vastly simplifies the rest of the pipeline. Carrying multiple sets of possible parameters complicates how properties such as the significance are calculated, so keeping just one set of source parameters is ideal.

Forming these coincidences drastically decreases the chance that any one trigger could be due to a glitch as it is far less likely that a loud noise transient occurs simultaneously across detectors. However, as will be discussed in the following section, the GstLAL pipeline does additionally allow for single-detector gravitational wave candidates. In this case, triggers in one detector which do not have a matching trigger in the others are still considered as possible events and the GstLAL pipeline applies penalty to their significance as discussed in the next section.

2.1.5 Ranking & Clustering

Despite the many mitigation plans put into place by GstLAL, glitches can still be present in the data and analyzed by the pipeline. In these situations, the SNR timeseries itself is not enough to distinguish between a real signal and a glitch as both could match well with templates in the bank. Therefore, the GstLAL pipeline also introduces a signal consistency check, ξ^2, which is a measure of how similar the observed SNR timeseries is to the predicted SNR timeseries of a signal exactly matching one of the templates [49]. This ideal timeseries can be calculated using the auto-correlation of the complex template waveform with the measured SNR. The complex autocorrelation function is then [49]:

$$R_i(t) = \int_{-\infty}^{\infty} \frac{|\tilde{h}_{+,i}|^2 + |\tilde{h}_{\times,i}|^2}{S_n(f)} e^{2\pi i f t} df \tag{2.26}$$

where i is the template index, and $t = 0$ is chosen to be the peak time, or the time of the peak SNR, and $R_j(0) = 1$. The signal consistency test ξ^2, is then defined as the amplitude squared of the difference between the observed SNR timeseries, and the predicted:

$$\xi_i^2(t) = |\rho_i(t) - \rho_i(0)R_i(t)|^2 \tag{2.27}$$

where ρ_i is the observed complex SNR timeseries given by equation 2.25. In the limit that the data is Gaussian noise, or $\tilde{d}(f) = \tilde{n}(f)$, the expectation value of the test, $< \xi^2 >= 2 - 2|R_i(t)|^2$. In practice, ξ^2 is calculated for some small window around the peak time of the event, typically about 86ms, and normalized as [49]:

$$\xi_i^2(t) = \frac{\int_{-\delta t}^{\delta t} |\rho_i(t) - \rho_i(0)R_i(t)|^2 dt}{\int_{\delta t}^{-\delta t} (2 - 2|R_i(t)|^2) dt} \tag{2.28}$$

This value is calculated at the time of trigger generation, and is stored along with the other trigger parameters.

Each trigger individually then has its SNR, ξ^2, and t_p calculated for each detector. Once coincidences are formed, or single detector candidates identified, they need to be ranked, and have a significance assigned. The GstLAL analysis uses the likelihood-ratio, as suggested by the Neyman-Pearson lemma, to rank events. We follow here the description of the likelihood ratio following [49], and it can be given generally as:

$$\mathcal{L} = \frac{P(\vec{D}_H, \vec{O}, \vec{\rho}, \vec{\xi^2}, \Delta\vec{\phi}, \Delta\vec{t}, \bar{\theta} \mid \text{signal})}{P(\vec{D}_H, \vec{O}, \vec{\rho}, \vec{\xi^2}, \Delta\vec{\phi}, \Delta\vec{t}, \bar{\theta} \mid \text{noise})} \tag{2.29}$$

where D_H is the horizon distance, $\vec{O}$ is the set of observing detectors, $\bar{\theta}$ is the label for the background bin of the template, and $\Delta\phi$ and Δt are the phase and time difference between any pair of detectors. $\vec{}$denotes a vector across all detectors e.g. if L1 and H1 are used, then $\vec{O} = \{O_{L1}, O_{H1}\}$

The denominator is a measure of how likely it is that the signal is not a true signal and instead is some kind of noise. It can be factored as follows:

$$P(...|n) = P(t_{ref}, \theta|n) \times P(\vec{O}|t_{ref}, \theta, n) \times P(\Delta\vec{t}, \vec{\phi}|\vec{O}, n) \times P(\vec{\rho}, \vec{\xi^2}|t_{ref}, \theta, n) \tag{2.30}$$

where $t_{ref} = \vec{t} - \vec{\Delta t}$ and is the time in the reference detector. $P(t_{ref}, \theta|n)$ is the probability that an event with parameters defined by θ occurs at time t_{ref} in any detector combination. It is estimated as the mean rate of coincident events in a single detector combination then summed across every possible detector combination. The mean rate in a single detector combination is found assuming Poisson statistics for noise events as the single detector rate for each detector in the combination multiplied together and times the coincidence window, typically 10ms. For example, for a coincidence in H1 and L1, the mean rate would be the rate in L1 times the rate in H1 times the coincidence window. Then $P(t_{ref}, \theta|n)$ would be the mean rate of coincidence in H1 and L1 plus the mean rate in H1 plus the mean rate in L1.

$P(\vec{O}|t_{ref}, \theta, n)$ is the mean rate of the coincident events in the detector combination which was operating at the time of the event weighted by the mean rate of all possible combinations given by $P(t_{ref}, \theta|n)$. $P(\vec{\Delta t}, \vec{\phi}|\vec{O}, n)$ is the probability of seeing the data with the specific time and phase difference between detectors, but for noise the time and phase distributions are approximated to be uniform and therefore this term is constant.

Finally, $P(\vec{\rho}, \vec{\xi^2}|t_{ref}, \theta, n)$, is the background distribution of the SNR and signal consistency across detectors for noise. This is generated by collecting the SNR and ξ^2 terms of single detector events during observing time with more than one detector, but not single detector events during observing time with just one detector. These histograms are constructed for the noise hypothesis, and so coincident events, and single detector events during single detector observing time, are not added to them as they are considered to be possible GWs. These samples are then populated into histograms, and grouped by their background bin, $\bar{\theta}$. The division of these background bins is described in 2.1.1. Each bin's histogram has a Gaussian KDE applied to make the likelihood ratio distribution smooth and a plot of a representative bin can be seen in Figure 2.3. These histograms are recorded on disk every four hours during a low-latency analysis, so that only up to four hours of data can be lost at any given time. Note that these histograms are a function of time for low-latency analyses as the histograms are populated as coincident detector data is available, and it accumulates over time. Low-latency analyses therefore require a "burn-in" time when they are first launched, while they accumulate a large enough sample in these histograms to be able to assign a likelihood ratio. The amount time required depends entirely on the up time of the detectors, but in a typical observing run takes $O(3)$ days.

The numerator meanwhile is measure of how likely it is that the data observed is a true signal, and not just noise. It factors as:

$$P(...|\text{signal}) = P(\theta|s) \times P(t_{ref}, \phi_{ref}|\theta, s) \times P(\vec{O}|t_{ref}, s) \times P(\vec{\rho}, \vec{\Delta t}, \vec{\phi}|\vec{O}, t_{ref}, s) \times P(\vec{\xi^2}|\vec{\rho}, \theta, s) \quad (2.31)$$

$P(\theta|s)$ constitutes the prior, and is a measure of how likely the template with parameters θ is to recover a gravitational wave. This information comes from the population model which contains information such as how possible sources are distributed in the template bank. $P(t_{ref}, \phi_{ref}|\theta, s)$ is taken to be $\propto D_H^3(t_{ref}, \theta)$ where D_H is the horizon distance. As discussed in 2.1.2, the PSD is calculated and updated continuously by the analysis, so the horizon distance for each detector is calculated from this value using a fiducial SNR of 8, and stored on disk for use in this term.

$P(\vec{O}|t_{ref}, s)$ is a similar term as in the noise model, except now it is the probability that the

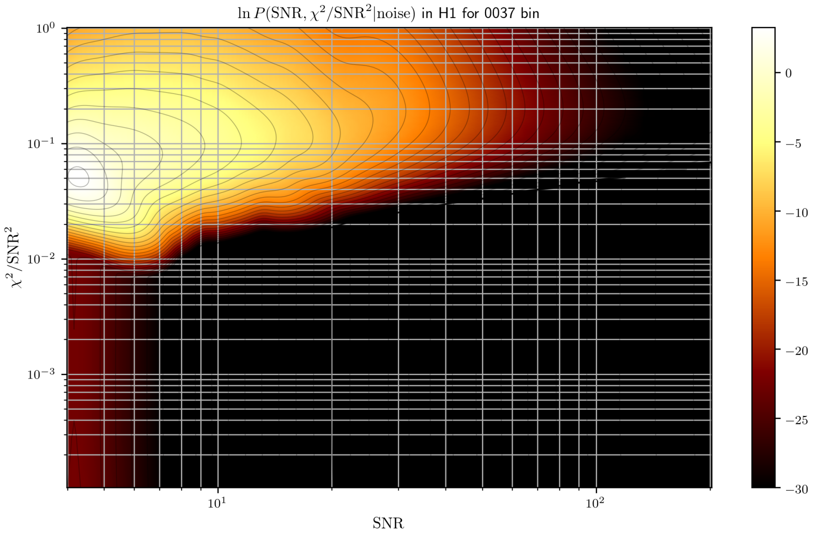


Figure 2.3. Plot of the SNR, χ^2 distribution for the noise model for the 37th SVD bin of Hanford. The y axis shows the χ^2 distribution, the x-axis is the SNR and the color bar is the total lnP(SNR,$\chi^2/$SNR^2|noise term in the likelihood ratio. There is a concentration of high values around high χ^2 and low-SNR, or the typical noise trigger parameter space. Real gravitational wave events would then lie outside of these high lnP areas in the dark regions of this figure.

data will appear in this set of detectors as a signal, e.g. above an SNR threshold. This is done by simulating GW sources distributed across the sky and at a variety of horizon distances, and performing a Monte Carlo integration to find detectable sources up to a horizon distance of interest. The details of this are given in [50]. $P(\vec{\rho}, \vec{\Delta t}, \vec{\phi}|\vec{O}, t_{ref}, s)$ gives the probability of having detected a signal with a given SNR, time difference, and phase shift between the observed detectors and acts as another signal consistency test. Details on how this term is calculated can be found in [49]. For single detector candidates, or those GW candidates detected when only a single detector is in observing mode, this term becomes a tunable penalty to the overall ratio. Ahead of O4, the term was tuned to be -13 based on participation in Mock Data Challenges as described in section 4.3.1.

Finally, $P(\vec{\xi^2}|\vec{\rho}, \theta, s)$ is again the previously mentioned ξ^2 signal consistency check, but this time under the assumption of a signal instead of noise.

2.1.6 Significance

With the likelihood-ratio of an event calculated, its true significance is also relevant. In other words, how often would an event with this likelihood ratio occur by chance with noise? If noise might generate the event frequently, then its less statistically significant say than an event that might only be generated once per hundred years by noise. This significance is estimated by tracking the denominator of the likelihood ratio, or the probability of an event occurring given that it is noise. The value of the entire denominator is recorded per noise event, where a noise event is defined as a single detector event during multiple detector observing time, as it was

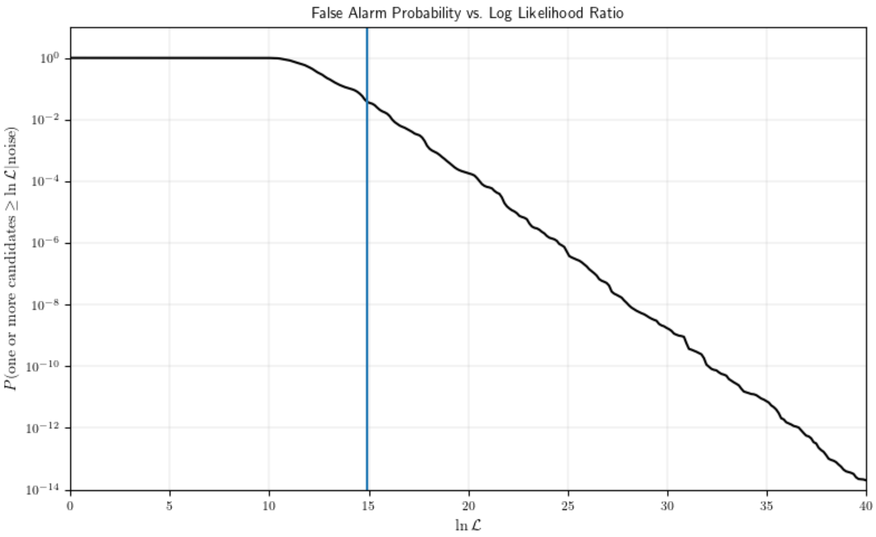


Figure 2.4. CCDF of a gravitational wave candidate. The x-axis denotes the value of the log-likelihood ratio of a given gravitational wave event and shows how that value is mapped to the false alarm probability on the y-axis.

for the $P(\vec{\rho}, \vec{\xi^2}|t_{ref}, \theta, n)$ term. These values are collected per background bin during a burn-in period for an online-analysis, or per relevant time segment in an offline analysis. Once there are enough samples in the histogram for a given background bin, the posterior density function (PDF) of this distribution is calculated. This is equivalent to integrating the individual terms in the denominator over constant likelihood surfaces, and both methods return the probability of a certain likelihood ratio value being generated by noise in a particular background bin, or $P(\mathcal{L}|\theta, \text{noise})$.

$P(\mathcal{L}|\theta, \text{noise})$ is then marginalized over all the background bins in the analysis returning a single $P(\mathcal{L}|\text{noise})$. The complimentary cumulative distribution (CCDF) of this PDF is then given as:

$$C(\mathcal{L}^*|n) = P(\mathcal{L} \geq \mathcal{L}^*|n) \int_{\mathcal{L}^*}^{\infty} P(\mathcal{L}|n) d\mathcal{L} \tag{2.32}$$

where $C(\mathcal{L}|n)$ gives the probability of a random noise event getting a likelihood ratio of $\mathcal{L}^*$ or higher. An example of such a CCDF can be seen in Figure 2.4 where the blue line denotes an example likelihood ratio of an event.

Dividing this by the total observing time up to this point in a low-latency analysis, or the total time during the entire run for an offline analysis, then gives the false alarm rate (FAR) or the rate at which events like this one would be generated by noise [23]:

$$FAR = \frac{C(\mathcal{L}|n)}{T} \tag{2.33}$$

where T is the observing time.

To account for observing multiple coincidences, one can calculate the false alarm probability using the binomial distribution as:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \tag{2.34}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{2.35}$$

where X is a random variable, n is the number of trials, k is the number of successes, and p is the rate. In this case, there would be M tests of detecting a noise event with an likelihood of $\mathcal{L}^*$ or higher where M is the number of gravitational wave candidates observed with that same likelihood threshold. For the false alarm probability, then the number of successes would need to be at least one, in other words, the probability that out of all M events at least one of them were generated by noise. This can be given as one minus the probability of none of them being generated by noise. In other words [23]:

$$\begin{align} P(\mathcal{L}^*|n_0, ..., n_M) &= 1 - \binom{M}{0} P(\mathcal{L}^*|n)^0 (1 - P(\mathcal{L}^*|n)^M \tag{2.36} \\ &= 1 - (1 - P(\mathcal{L}^*|n)^M \tag{2.37} \\ & \tag{2.38} \end{align}$$

where the second equation is given by the binomial and following term being identically one. This value is then updated as more events are detected during a run, whereas the FAR only takes into account the current event and the noise background without incorporating any knowledge of other signal events.

2.2 Workflow

The methodology introduced in Chapter 2, will be discussed in the context of both online, or low-latency, and offline, or high-latency, filtering here. A comparison of the O4 online and offline sensitivities is discussed in detail in Chapter 5.

2.2.1 Online

The online analysis is by definition required to be low-latency and keep up with real-time data availability. Therefore, it can only use historical data during the ranking of candidates, and does not have the benefit that we will see in the offline analysis of using data from future times in its processing. However, as discussed in section 1.1, the crucial role that the prompt discovery of gravitational wave events plays in the multi-messenger astronomy ecosystem cannot be understated and a potential loss in sensitivity is well worth the scientific gain of low-latency detection.

The template bank, its LLOID decomposition, and a PSD to seed the online analysis are all pre-computed and supplied as inputs to the low-latency search. The whitening, conditioning, filtering, background collection, and trigger generation are all completed in parallel processes per background bin. Then, a separate process is used to periodically marginalize over the backgrounds collected by each bin, and combine them for use in FAR assignment. At the time of a gravitational wave candidate, then, the log-likelihood and FAR are assigned to the event based on only the background information available at the time to the analysis, while in offline this is not the case. This fact has been mitigated during O4 with the implementation of a re-ranking procedure described in Chapter 5.

As many bins may all identify triggers for a single gravitational wave event, an additional process is required to aggregate these triggers. Only the highest SNR of all triggers in a small time window (about 0.5 seconds) is uploaded to the Gravitational wave Candidate Event Database (GraceDB) [14], a interactive database used to aggregate information across detection pipelines in real-time. There are additionally a handful of support processes which create and upload plots to GraceDB, collect monitoring, and more.

As a new feature in O4, streams of simulated data are handled by the online analysis in a similar manner. They follow the same processes as real data, but instead of collecting their own background, which would be contaminated by the rate of simulated signals, they instead take a copy of background data calculated by the same bin process which analyzes the real data. This keeps ranking of the simulated data most similar to the real data, allowing for an accurate estimate of the sensitivity of the online analysis. How this simulated data is supplied to the online analysis, and the features surrounding it are discussed in detail in Chapter 5.

2.2.2 Offline

The offline analysis has the benefit of foresight when it comes to ranking gravitational waves, and therefore can more accurately assign significance to gravitational wave events. The offline analysis completes each phase of its workflow in stages and is supplied with a list of segments defining the science quality data from each detector. It starts by calculating the running average PSD for each of these segments, called a reference PSD. This output can then be used as inputs to any number of pipelines analyzing the same data, e.g. a single observing run. For example, there may be many analyses completed which each target a different region of the parameter space, and they can all use this PSD as inputs.

Any offline analysis then takes the median of the reference PSDs to whiten the input template bank and generate the full decomposition of the bank following the LLOID method described in 2.1.1. This bank is then used as in the online case and supplied to a distributed number of processes which whiten, condition, and filter the data.

However, in the offline case, this data is not ranked by the processes which filter it. Instead, this is left to a second phase of processes which take a mass model, the filtered data, and a handful of other data products to produce triggers and rank it. This again is for agility, allowing for the mass model, for example, to be swapped out and a different ranking applied without being forced to use the computing resources to completely re-filter the data itself. During this ranking stage, the offline analysis takes background information collected across the entire observing run to rank candidates, supplying it with a more informed model and resulting in a more sensitive search.

Simulated data is again treated in a similar method to the online case. During the ranking stage, it reads the background collected from the normal filtered data and uses it to rank both the injection and non-injection data.

Chapter 3
Offline Results from LIGO and Virgo's Third Observing Run

The LIGO and Virgo Collaboration's third observing (O3) run ran from April 2019 to March 2020, and was split into two halves O3a, and O3b. The latter half of this observing run, O3b, ran from November 1, 2019 to March 27, 2020 with the LIGO Hanford (LHO), LIGO Livingson (LLO), and Virgo interferometers participating. I contributed to the operation of the high-latency, offline GstLAL search covering O3b and the results from this half are summarized here, but full reporting of O3 results can be found at [51], [52], and [13].

We will start with a short description of the upgrades implemented at the LHO and LLO interferometers ahead of O3, and between O3a and O3b. We then discuss the gravitational wave candidates found during O3b, highlighting exceptional events and the role GstLAL played in their discovery. Finally, we summarize the impact these events have on the known rates and populations of compact binary sources in our universe.

3.1 O3 Detector Upgrades

There were several detector upgrades to the LIGO Hanford and LIGO Livingston interferometers which are discussed in detail in [13], but are summarized here.

At both LHO and LLO, improvements to the mirror systems were made to reduce radiation pressure while increasing laser input power. The mirrors which reflect light back to the beam splitter from the arm stations were first replaced. The replacement mirrors had lower scattering losses, thereby maintaining more laser power after every incidence. An increase in laser power, however, increases the radiation pressure on the mirrors thereby inducing excitation in the mirror mechanical modes and increasing the noise. Therefore, ahead of O3a, acoustic mode dampers were installed at both sites [53]. The acoustic mode dampers attach to the mirrors with springs to passively damped the excitation modes induced by the radiation pressure. This allows for an increase in the laser power while mitigating the increase in radiation pressure noise at low frequency.

The light sources themselves were also renovated ahead of O3 with higher input power lasers, and the implementation of squeezed light sources. The base state of the detector has equal uncertainty in the phase and amplitude (e.g. position and momentum) of the light incident on the output optics. This results in equal contributions from the photon shot noise and radiation pressure noise described in section 1.3.2. However, in a squeezed state, the uncertainty in one is decreased at the expense of increased uncertainty in the other. In this case, the photon shot noise was reduced with an increase in the radiation pressure noise thereby lowering the noise floor at high frequency, and increasing it at low frequency. Between O3a and O3b, even higher sensitivity was achieved at both sites by implementing further improvements to the crystal and tuning of the system used for squeezing.

Each of the detector sites additionally experienced noise sources unique to their locations. At LLO, seismic activity created by ocean waves caused the distance between the main mirrors, and a secondary reaction mirror to vary. This variation caused portions of the laser light to scatter outside of its intended path by changing the incidence. This scattered laser light can then reflect off of another surface and back into the main beam with a different phase, thereby causing scattering noise. Scattering noise causes Scattered Light glitches by resonating harmonics of the surface motion frequencies of the mirrors. Therefore, at both detectors, additional hardware was added which prevents the scattered light from re-entering the main beam. The abundance of this noise at LLO additionally prompted the creation of a new control linking the movement of the reaction mirror to that of the main mirror during high seismic times. Both of these improvements were seen to significantly reduce the number of Scattered Light glitches present between O3a and O3b.

3.2 Results

35 new candidates were discovered during O3b, bringing the total number of candidates up to this point to 90. Inclusion as a candidate gravitational wave is based on FAR and p_{astro} thresholds which vary for low-latency alerts and offline candidacy. As described in section 2.1.6, FAR is an estimate of how often the candidate may be caused by random noise fluctuations in the detectors, and is therefore an immediate estimate of the significance of an event. However, this is highly dependent on the noise events collected by each detection pipeline, and how these events are collected vary pipeline to pipeline. Additionally, FAR does not take into account any of our previous knowledge of astrophysical distributions, or of detections made thus far. p_{astro}, meanwhile, accounts for the probability that an event is of astrophysical origin given what we know about likely source distributions, and can be informed by previous detections.

For example, an event with a low FAR may appear to come from a source class that has never been detected before, and which is extremely unlikely given what we previously know about its properties. This would make this event have a lower p_{astro}, despite its high FAR. Vice-versa an event may have a relatively low FAR, but if it originates from a source population which is known to frequently generate gravitational waves, then its p_{astro} would indicate it as more significant than its FAR alone. Therefore p_{astro} is useful for comparing events which could be from any number of possible sources, and represents a more complete estimate of the probability that a candidate is a real event given its source parameters.

In the low-latency analyses, a strict FAR threshold of 1.2/yr is imposed on alerts so as only to trigger follow-up from astronomers in the most confident cases. There is no threshold on p_{astro} because of a number of technical reasons, and the full accounting of a candidate's source parameters has a high uncertainty in low-latency treatment. This is not the case with offline analysis, however, and full offline candidates therefore must must instead meet a $p_{\text{astro}} > 0.5$ limit. If an event does not meet the p_{astro} limit offline, but does have a FAR of less than 2 per year, then it is considered a marginal candidate.

During low-latency, 39 events were reported at a FAR threshold of 1.2/year, but after further analysis only 18 are considered as offline candidates. Of the lost candidates, 16 were retracted, meaning that they are believed to be of terrestrial origin. The remaining 5 simply did not meet the threshold for inclusion as candidates following the offline thresholds. In the offline analyses, 17 new candidate events were found which did not pass the 1.2/yr FAR threshold in the online searches. All of them are coincident events between at least two detectors, and the majority fall into the BBH mass regions.

There are a couple reasons that events such as these may be considered as identified offline and not in low-latency. The data quality is better for offline data which has had the chance to be more closely cleaned and calibrated than in low-latency, and detection pipelines are provided with data quality vetoes to remove particularly noisy data. Additionally, as discussed in section 2.2, low-latency pipelines can only use historical data at the time of ranking an event, while offline pipelines have the benefit of knowing a priori about data throughout the entire observation period. This often results in improved background estimates, noise rejection, and significance estimates. For a discussion on improvements to the GstLAL search to account for these differences, see Chapter 5.

3.2.1 Gravitational-Wave Candidate List

Table 3.1 contains all of the candidates across offline pipelines for O3b and contains all 35 events considered as gravitational wave candidates during this period. The FAR, SNR, and p_{astro} from each pipeline is reported individually along with the active instruments following the Hanford(H), Livingston(L), and Virgo(V) convention and the event title. The candidate titles follow a date naming scheme with the year, month, day followed by an underscore with the hour, minute, second UTC designation. Additionally, those in bold were the 17 candidates not identified in low-latency, but which were identified for the first time offline. The italics signify when the italicized pipeline identified it below the threshold for inclusion, but another pipeline identified it above that threshold.

Five gravitational wave detection pipelines analyzed data using their offline configurations: cWB [54], GstLAL, MBTA [55], PyCBC [56]. PyCBC additionally operated with two different configurations, noted as PyCBC-broad and PyCBC-BBH. The SPIIR pipeline [57] participated in low-latency detections, but did not re-filter data in an offline configuration, and is therefore not included in the discussion here. GstLAL, MBTA, and PyCBC use matching filtering with waveform templates to identify candidates, while cWB searches for un-modeled signals. Additionally, cWB only searches for candidates in detector pairs, while PyCBC and MBTA look for coincidences in two and three detectors, and GstLAL searches for signals in any detector combination. GstLAL compares possible candidates to the full background from across O3b, while the other three pipelines use background from more locally around a candidate. These differences, and others, result in differences in the estimation of p_{astro} and FAR, and therefore pipelines are expected to detect a slightly different subset of events, particularly at low SNR.

These differences can be seen in practice in the results shown in Table 3.1. Of all 35 candidates, 21 were found by more than one pipeline while only 9 were identified by all of them. 29 candidates were found by one or both of the PyCBC searches, while GstLAL identified 21, MBTA found 20, and 10 candidates were recovered by cWB. Additionally, of these candidates, many were found only by a single pipeline. MBTA and PyCBC contributed four, and eight candidates unique to their pipelines respectively. The MBTA candidates are signals from BBH sources, and are included in the candidate list because they have high p_{astro} despite also having high FAR estimates. Of the 8 unique PyCBC events, the most notable is the potential NSBH GW191219_163120, which is described further in the next section. The 7 other unique candidates are from BBH sources, and are generally middling to high FAR with the highest FAR among them given as 0.46 per year.

GstLAL identified 2 unique candidates. Both of these unique events were from single detector triggers, or an event found significantly in only one detector when one or more were operating. In these cases, they were found when Livingston was operating alongside Virgo. Single detector events in a network of HV or LV are not implausible as the detector sensitivity of Hanford and Livingston are both approximately double that of Virgo. However, GstLAL is the only pipeline which searched for these events during O3. Other pipelines do not consider these times for possible candidate identification to prevent against the likelihood of false detection claims, but the two events GstLAL recovered during this time provide strong evidence for the analysis of these times. The first unique event, GW200112_155838 is a BBH event with masses in the $26-42M_{\odot}$ range, with a low spin, but within a redshift of 0.25. The second event, GW200105_162426 is discussed in more detail in 3.2.1.1 as it is a potential NSBH event.

While differences in sensitivities and configurations creates so many unique pipeline candidates, there are many more which are discovered in a subset combination of the four pipelines for the same reasons. However, GW200129_0654580 is unique among them as it was found only by GstLAL and PyCBC because of a data quality issue and not because of general differences in pipeline sensitivity. During the time surrounding GW200129_0654580, a Livingston data issue was flagged, and included in a CAT-2 veto set (described in 1.4) distributed to search pipelines. Implementation of this veto set caused PyCBC, MBTA, and cWB to ignore Livingston data around the time of the event. This caused MBTA and cWB to miss this event, while PyCBC recovered it using just the HV data, but lost SNR compared to the result from HLV.

GstLAL, however, did not implement CAT-2 vetoes at the time, and its worth noting that CAT-2 vetoes are not being produced by the collaboration during O4. GstLAL depends on the pipeline including noise events in the background in order to have an accurate ranking, and therefore analyzes data normally flagged by CAT-2 vetoes to make the background more robust. Additionally, GstLAL was using an implementation of iDQ at the time which down-ranked events based on the data quality it reported [58], and therefore relied on it over the removal of times via CAT-2 vetoes. This allowed GstLAL to analyze the full set of HLV information, and demonstrates the power of incorporating additional data quality metrics into search pipelines.

Table 3.2 shows the marginal candidates, or those with FAR below 2.0 per year, but a p_{astro} above the threshold of 0.5 for inclusion in the main candidate table. Those in this table without the GW prefix, 200121_031748, 200214_224526 and 200219_201407, were found

Name	Inst.	CWB			GstLAL			MBTA			PyCBC-broad			PyCBC-BBH		
		FAR (yr^{-1})	SNR	p_{astro}	FAR (yr^{-1})	SNR	p_{astro}	FAR (yr^{-1})	SNR	p_{astro}	FAR (yr^{-1})	SNR	p_{astro}	FAR (yr^{-1})	SNR	p_{astro}
GW191103_012549	HL	–	–	–	–	–	–	*27*	*9.0*	*0.13*	4.8	9.3	0.77	0.46	9.3	0.94
GW191105_143521	HLV	–	–	–	*24*	*10.0*	*0.07*	0.14	10.7	> 0.99	0.012	9.8	> 0.99	0.036	9.8	> 0.99
GW191109_010717	HL	< 0.0011	15.6	> 0.99	0.0010	15.8	> 0.99	1.8×10^{-4}	15.2	> 0.99	0.096	13.2	> 0.99	0.047	14.4	> 0.99
GW191113_071753	HLV	–	–	–	–	–	–	26	9.2	0.68	*1.1×10^{4}*	*8.3*	*< 0.01*	*1.2×10^{3}*	*8.5*	*< 0.01*
GW191126_115259	HL	–	–	–	*80*	*8.7*	*0.02*	*59*	*8.5*	*0.30*	*22*	*8.5*	*0.39*	3.2	8.5	0.70
GW191127_050227	HLV	–	–	–	*0.25*	*10.3*	*0.49*	1.2	9.8	0.73	*20*	*9.5*	*0.47*	4.1	8.7	0.74
GW191129_134029	HL	–	–	–	$< 1.0 \times 10^{-5}$	13.3	> 0.99	0.013	12.7	> 0.99	$< 2.6 \times 10^{-5}$	12.9	> 0.99	$< 2.4 \times 10^{-5}$	12.9	> 0.99
GW191204_110529	HL	–	–	–	*21*	*9.0*	*0.07*	*1.3×10^{4}*	*8.1*	*< 0.01*	*980*	*8.9*	*< 0.01*	3.3	8.9	0.74
GW191204_171526	HL	$< 8.7 \times 10^{-4}$	17.1	> 0.99	$< 1.0 \times 10^{-5}$	15.6	> 0.99	$< 1.0 \times 10^{-5}$	17.1	> 0.99	$< 1.4 \times 10^{-5}$	16.9	> 0.99	$< 1.2 \times 10^{-5}$	16.9	> 0.99
GW191215_223052	HLV	0.12	9.8	0.95	$< 1.0 \times 10^{-5}$	10.9	> 0.99	0.22	10.8	> 0.99	0.0016	10.3	> 0.99	0.28	10.2	> 0.99
GW191216_213338	HV	–	–	–	$< 1.0 \times 10^{-5}$	18.6	> 0.99	9.3×10^{-4}	17.9	> 0.99	0.0019	18.3	> 0.99	7.6×10^{-4}	18.3	> 0.99
GW191219_163120	HLV	–	–	–	–	–	–	–	–	–	4.0	8.9	0.82	–	–	–
GW191222_033537	HL	$< 8.9 \times 10^{-4}$	11.1	> 0.99	$< 1.0 \times 10^{-5}$	12.0	> 0.99	0.0099	10.8	> 0.99	0.0021	11.5	> 0.99	9.8×10^{-5}	11.5	> 0.99
GW191230_180458	HLV	0.050	10.3	0.95	0.13	10.3	0.87	*8.1*	*9.8*	*0.40*	*52*	*9.6*	*0.29*	0.42	9.9	0.96
GW200112_155838	LV	–	–	–	$< 1.0 \times 10^{-5}$†	17.6	> 0.99	–	–	–	–	–	–	–	–	–
GW200115_042309	HLV	–	–	–	$< 1.0 \times 10^{-5}$	11.5	> 0.99	0.0055	11.2	> 0.99	$< 1.2 \times 10^{-4}$	10.8	> 0.99	–	–	–
GW200128_022011	HL	1.3	8.8	0.63	0.022	10.1	0.97	3.3	9.4	0.98	0.63	9.8	0.95	0.0043	9.9	> 0.99
GW200129_065458	HLV	–	–	–	$< 1.0 \times 10^{-5}$	26.5	> 0.99	–	–	–	$< 2.3 \times 10^{-5}$	16.3	> 0.99	$< 1.7 \times 10^{-5}$	16.2	> 0.99
GW200202_154313	HLV	–	–	–	$< 1.0 \times 10^{-5}$	11.3	> 0.99	–	–	–	–	–	–	0.025	10.8	> 0.99
GW200208_130117	HLV	–	–	–	0.0096	10.7	0.99	0.46	10.4	> 0.99	0.18	9.6	0.98	3.1×10^{-4}	10.8	> 0.99
GW200208_222617	HLV	–	–	–	*160*	*8.2*	*< 0.01*	*420*	*8.9*	*0.02*	–	–	–	4.8	7.9	0.70
GW200209_085452	HLV	–	–	–	0.046	10.0	0.95	12	9.7	0.97	*550*	*9.2*	*0.04*	1.2	9.2	0.89
GW200210_092254	HLV	–	–	–	*1.2*	*9.5*	*0.42*	–	–	–	17	8.9	0.53	7.7	8.9	0.54
GW200216_220804	HLV	–	–	–	0.35	9.4	0.77	*2.4×10^{3}*	*8.8*	*0.02*	*970*	*9.0*	*< 0.01*	7.8	8.7	0.54
GW200219_094415	HLV	0.77	9.7	0.85	9.9×10^{-4}	10.7	> 0.99	0.18	10.6	> 0.99	1.7	9.9	0.89	0.016	10.0	> 0.99
GW200220_061928	HLV	–	–	–	–	–	–	–	–	–	–	–	–	6.8	7.5	0.62
GW200220_124850	HL	–	–	–	*150*	*8.2*	*< 0.01*	1.8×10^{3}	8.2	0.83	–	–	–	*30*	*7.8*	*0.20*
GW200224_222234	HLV	$< 8.8 \times 10^{-4}$	18.8	> 0.99	$< 1.0 \times 10^{-5}$	18.9	> 0.99	$< 1.0 \times 10^{-5}$	19.0	> 0.99	$< 8.2 \times 10^{-5}$	19.2	> 0.99	$< 7.7 \times 10^{-5}$	18.6	> 0.99
GW200225_060421	HL	$< 8.8 \times 10^{-4}$	13.1	> 0.99	0.079	12.9	0.93	0.0049	12.5	> 0.99	$< 1.1 \times 10^{-5}$	12.3	> 0.99	4.1×10^{-5}	12.3	> 0.99
GW200302_015811	HV	–	–	–	0.11†	10.6	0.91	–	–	–	–	–	–	–	–	–
GW200306_093714	HL	–	–	–	–	–	–	410	8.5	0.81	*3.4×10^{3}*	*7.8*	*< 0.01*	*24*	*8.0*	*0.24*
GW200308_173609	HLV	–	–	–	*680*	*8.1*	*< 0.01*	*6.9×10^{4}*	*8.3*	*0.24*	*770*	*7.9*	*< 0.01*	2.4	8.0	0.86
GW200311_115853	HLV	$< 8.2 \times 10^{-4}$	16.2	> 0.99	$< 1.0 \times 10^{-5}$	17.7	> 0.99	$< 1.0 \times 10^{-5}$	16.5	> 0.99	$< 6.9 \times 10^{-5}$	17.0	> 0.99	$< 7.7 \times 10^{-5}$	17.4	> 0.99
GW200316_215756	HLV	–	–	–	$< 1.0 \times 10^{-5}$	10.1	> 0.99	*12*	*9.5*	*0.30*	0.20	9.3	0.98	0.58	9.3	0.98
GW200322_091133	HLV	–	–	–	–	–	–	450	9.0	0.62	*1.4×10^{3}*	*8.0*	*< 0.01*	*140*	*7.7*	*0.08*

Table 3.1. Candidate GW signals over the course of O3b. The time (UTC) of the signal is encoded in the name as GWYYMMDD_hhmmss (e.g., GW200112_155838 occurred on 2020-01-12 at 15:58:38). The names of candidates not previously reported by the low-latency analyses are given in **bold**. The detectors observing at the merger time of the candidate are indicated using single-letter identifiers (e.g., H for LIGO Hanford); these are not necessarily the same detectors that contributed triggers associated with the candidate. Where a candidate was found with p_{astro} above the threshold value of 0.5 by at least one analysis but below the threshold by others, we include in *italics* the results from the other analyses, where available. A dash (–) indicates that a candidate was not found by an analysis. The 2 candidates labeled with a dagger (†) were found only above threshold in a single detector with the GstLAL analysis, and the FAR estimates were made using significant extrapolation of the background data, meaning that single-detector candidates have higher uncertainty than coincident candidates. Reprinted from [13].

Name	Inst.	CWB			GstLAL			MBTA			PyCBC-broad		
		FAR (yr^{-1})	SNR	p_{astro}	FAR (yr^{-1})	SNR	p_{astro}	FAR (yr^{-1})	SNR	p_{astro}	FAR (yr^{-1})	SNR	p_{astro}
GW191118_212859	LV	–	–	–	–	–	–	*7.4×10^5*	*8.0*	*< 0.01*	1.3	9.1	0.05
GW200105_162426	LV	–	–	–	$0.20^{\dagger}$	13.9	0.36	–	–	–	–	–	–
200121_031748*	HV	–	–	–	*58*	*9.1*	*0.02*	1.1	10.7	0.23	–	–	–
GW200201_203549	HLV	–	–	–	1.4	9.0	0.12	*850*	*8.9*	*< 0.01*	*1.0×10^3*	*8.3*	*< 0.01*
200214_224526*	HLV	0.13	13.1	0.91	–	–	–	–	–	–	–	–	–
200219_201407*	HLV	–	–	–	–	–	–	0.22	13.6	0.48	–	–	–
GW200311_103121	HL	–	–	–	*110*	*9.0*	*< 0.01*	1.3	9.0	0.03	1.3	9.2	0.19

Table 3.2. Marginal candidates found by the various analyses. The candidates in this table have a FAR below a threshold of 2.0 yr^{-1} in at least one analysis, but were not found with p_{astro} that meets the threshold for Table 3.1. Detector-identifying letters are the same as given in Table 3.1. The instruments for each candidate are the ones which were operating at the time of the trigger, and are not necessarily the same as those which participated in the detection. The candidates are named according to the same convention as in Table 3.1 except that here we omit the GW prefix for the candidates found to be likely caused by instrumental artifacts, indicated with an asterisk ($*$). Where a candidate was seen below the FAR threshold in at least one analysis but above threshold in others, we include in *italics* the information on that trigger from the other analyses as well where available. As in Table 3.1, the dagger (†) indicates a candidate found by a single detector with the GstLAL analysis. Reprinted from [13]

to be from instrumental artifacts. During all three events, one or both of the detectors were known to contain loud glitches during or around the trigger time. During 200121_031748, LHO contained a loud blip glitch and this glitch was recovered by GstLAL and PyCBC at a FAR of 58 and 1.1, respectively. The low FAR from PyCBC caused this event to be classified as a marginal event and highlights the uncertainty of the events in this category. 200214_224526 was recovered only by cWB, and during this event LLO and LHO both contained fast scattering glitches. MBTA alone recovered 200219_201407, and LHO had many loud glitches present in a small range around the event time.

Of the remaining four marginal candidates, a subset of three of them were recovered by PyCBC, MBTA, and GstLAL. As a potential NSBH detection during single detector time by GstLAL, GW200105_62426 will be discussed further in the next section. GW200311_103121 is a potential BNS event recovered significantly by MBTA and PyCBC at a FAR of 1.3 per year, and with a less significant trigger from GstLAL with a FAR of 110 per year. If real, this event has a template consistent with a chirp mass of only $1.17M_{\odot}$ making it an extremely light system. However, its p_{astro} of less than 20% from either pipeline make it a marginal event, and only more observations will better inform the origin of this candidate. The other two events GW191118_212859 and GW200201_203549 have templates consistent with a BBH source, but are only included here because of the significance reported in a single pipeline and full parameter estimation has not been done on these events.

There are additionally 1041 more events which do not pass the p_{astro} or FAR thresholds for inclusion in either the candidate or marginal event table, but do pass a FAR threshold of 2 per day. These sub-thresholds events have generally not been followed-up on, or examined further by the collaboration, but are used in some rate estimates as discussed further in the next section.

3.2.1.1 Potential NSBH Events

Of the 35 candidates in O3b, GW191219_163120, GW200115_042309, GW200210_092254 and GW200105_162426 were the only which were consistent with a NSBH system, that is with one mass less than three solar masses. As introduced in section 1.2.2, the shape of a neutron star can be changed in the presence of tidal forces from its companion in a binary. How easily the neutron star is deformed is governed by its tidal deformability, Λ, and its value is imprinted on gravitational waveforms at fifth PN order. Therefore, it may be possible to determine if any GW binary contains a neutron star simply by measuring its tidal deformability as imprinted on the waveform. Unfortunately, at current detector sensitivities the SNR of any event containing a neutron star would have to be quite large in order to constrain the tidal deformability since it enters at such a high order. This has been the case for GW170817 [17], and continues to be so here. Therefore, these events are considered to be NSBH solely by their inferred mass ranges.

GW200115_042309 in particular has the lowest total mass of all binaries in O3b. Its source masses are estimated at $m_1 = 5.9^{2.0}_{-2.5} M_\odot$ and $m_2 = 1.44^{0.85}_{-0.28}$, and a chrip mass of $\mathcal{M}_c = 2.43^{0.05}_{-0.07} M_\odot$ [13]. At these masses, the primary is consistent with a black hole while the secondary is consistent with a neutron star. The black hole primary additionally has a 29% probability of having a smaller than $5M_\odot$ mass [59] where $5M_\odot$ is the theorized lower black hole mass gap [60]. However, despite the 31 GCN Circulars reporting follow-up observations for this event, none of them, to date, claim an EM counterpart [59].

GW200105_162426 is a possible NSBH candidate, detected only by GstLAL as a single detector candidate with a FAR of 0.2/yr. Its primary masses are quoted at just $m_1 = 9.1^{1.7}_{-1.7} M_\odot$ and $m_2 = 1.91^{0.33}_{-0.24} M_\odot$ [59], making it the second-lightest total mass system in the catalogue. Its p_{astro} estimate, however, causes it to appear in the marginal event table instead of the full candidate list. At the time of detection, Livingston and Virgo were operating, but the signal was only recovered in Livingston. While PyCBC, MBTA, and GstLAL all have triggers from Livingston at that time, only GstLAL considered it as a candidate because of its single-detector status. However, both in real-time and offline GstLAL recovered it with a significant FAR of 0.36/yr and 0.2/yr respectively. The event is a clear departure from the background noise, and is distinct from all other noise events across the entirety of O3 – including other single detector events [59]. As there has only been $\mathcal{O}(1)$ NSBH event up to this point, it is possible that the uncertainties in the p_{astro} estimate cause it to falsely fall below the threshold for full inclusion. Unfortunately, there is no way to currently know whether this event will ever be promoted to a full candidate as improved merger rates and more accurate population models can only come from more detections. Additionally, of the 21 GCN circulars for this event, no EM counterpart has been reported [59].

GW200210_092254 seems to be another potential NSBH, but is thought to be more consistent with a high mass ratio BBH. Its primary source mass is easily consistent with a black hole at $m_1 = 24.1^{7.5}_{-4.6} M_\odot$, while its secondary mass is of more uncertain classification at $m_2 = 2.83^{0.47}_{-0.42} M_\odot$. The secondary mass has a probability of 76% for being less than $3M_\odot$, and therefore sits above the hypothesized maximum NS mass of approximately $2M_\odot$ [61], but below the lower mass bound on black holes [60]. This is in similarity to GW190814, whose component masses followed a similar pattern. However, given the more stringent bound on the upper mass limit now known for neutron stars, it is more likely that the the secondary mass is an extremely light black hole.

GW191219_163120 has a secondary mass which is the lightest in the catalogue. Its primary masses are inferred as $m_1 = 31.1^{+2.2}_{-2.8} M_\odot$ and $m_2 = 1.17^{+0.07}_{-0.06} M_\odot$. This secondary mass is so light that it is consistent with some of the lightest known neutron stars and makes the mass ratio of the binary at $q = 0.038^{+0.005}_{-0.004}$ so extreme that it is difficult to model. Otherwise, its primary mass is consistent with a black hole. This candidate was recovered only by the PyCBC-broad analysis at a FAR of 4 per year. As this FAR does not meet the traditional threshold, it is not included when calculating astrophysical implications in many analyses such as the ones done for GW200105_162426 and GW200115_042309 in [59].

This leaves two significant and confident NSBH events. Previously, there were only two potential NSBH events recorded with signficance: GW190814, and GW190426. However, GW190814 is now believed to be a BBH with a significantly light secondary mass and no longer consistent with a NS. These two events then contribute significantly to the total number of NSBH events observed to date by gravitational waves, and this has impacts on the rate and populations of both compact binaries of this source type.

3.3 Impact on Merger Rates and Mass Populations

The additional 35 events discovered during O3b have had an impact on the estimated rates and populations of CBC sources – particularly those with asymmetric mass ratios. Here, we summarize the results of [6], but a full treatment can be seen there. We classify three populations by dividing the mass parameter space. NS masses lie between $1 - 2.5 M_\odot$, and BH between $2.5 - 100 M_\odot$ where a BNS has two NS components, an NSBH one NS and BH, and a BBH has two BH. Note that this division is slightly different than the one used to label p_{astro} e.g. in the previous section where the NS range was $1 - 3 M_\odot$. The merger rates are calculated as:

$$\mathcal{R}(z) = \frac{dN}{dV_c dt}(z) = \mathcal{R}_0 (1+z)^\kappa \tag{3.1}$$

where R_0 is the local merger rate density at z=0, and κ is the parameter which controls the evolution of $\mathcal{R}$ at higher redshift. Generally, because of the sensitivity of our detectors, most BNS and NSBH detections only go out to a modest redshift, and therefore rates are calculated at $\kappa = 0$. However, for BBH populations this is not the case and so merger rates must be discussed as both a function of mass and redshift.

With these classifications, the BNS merger rate including the new GWTC-3 data is given as $105.5^{+190.2}_{-83.9} Gpc^{-3} yr^{-1}$. In this calculation, it was assumed that the merger rate is constant out to z=0.15, and spin magnitudes are less than 0.4. This is only about a third the rate of the previously quoted $320^{+490}_{-240} Gpc^{-3} yr^{-1}$ which considered data through GWTC-2 [62]. However, as there were no BNS detections during this period, this is to be expected.

For the NSBH population, we discuss a few different methods for a rate calculation following [59], constituting the first direct measurements of the NSBH merger rate. First, we only consider the two confident NSBH events GW200105 and GW200115 and treat each as one Poisson distributed count in the observing time up to the end of O3b. The two event-based rates for these events are then given as $\mathcal{R}_{200105} = 16^{+38}_{-14} Gpc^{-3} yr^{-1}$, and $\mathcal{R}_{200115} = 36^{+82}_{-30} Gpc^{-3} yr^{-1}$

assuming an uninformative prior. Combining the two, the total event-based rate is then $\mathcal{R}_{\mathrm{NSBH}} = 45^{+75}_{-33} Gpc^{-3} yr^{-1}$.

Second, we consider not just these two events but additionally all less significant triggers who fall into the mass ranges of $m_1 \in [2.5, 40] M_{\odot}$ and $m_2 \in [1, 3] M_{\odot}$, such as GW191219. Then, using the population of GstLAL triggers which fall in that range, the VT is calculated for the NSBH region, and the joint likelihood on Poisson parameters in the NSBH region, Λ_{NSBH}. The rate is then given by the joint likelihood over the VT as $\mathcal{R}_{\mathrm{NSBH}} = 130^{+112}_{-69} Gpc^{-3} yr^{-1}$. This rate is higher than the event-based rate, but includes more events as NSBH detections than the event-based rate does and events like GW190814 fall into this region.

Additionally, the mass population of NS specifically in binaries can be calculated using events up to the end of O3b. The mass distribution for these NS shows a flatter shape with more broad support up to a higher mass than the double-peaked Galactic NS distribution. Additionally, the maximum mass in the population is around two solar masses, while the galactic distribution supports up to 2.2 solar masses. However, the error in this measurement remains large, and more detections are needed to reduce the errors.

The excess of large mass ratio events in this catalog contribute significantly to the merger rate for masses between $1 - 10 M_{\odot}$. They cause a drop off in the event-based rate after BNS-like masses, with a further peak around high mass ratios consistent with NSBH masses, and then a final peak around equal mass BBH events. This structure can be seen clearly in Figure 3.1 for three models used in this analysis: Mulit-Source (MS), Power-Dip-Break (PDB) [63], and Binned-Gaussian-Process(BGP) [64] as well as a Fisher Matrix (FM) implementation for comparison. Because the NSBH rate occurs at a rate higher than BBH, but lower than BNS, it increases the rate of objects in the low mass BH region. This makes the event-based rates for mergers with one component in the mass gap between $2.5 - 5 M_{\odot}$ constrained to be less than that of the BNS merger rate, but consistent with the BBH rate.

Additionally, the BBH rate as a function of redshift and mass has been affected by the detections in this catalogue. First, the sub-structure in the merger rate as a function of primary mass posited in GWTC-2 [51] has been confirmed by the clustered mass of the BBH events discovered in O3b. This shows an over-density in the merger rate at primary masses around $m_1 = 10^{+0.29}_{-0.59} M_{\odot}$ and again, although with less intensity, at $m_1 = 35^{+1.7}_{-2.9} M_{\odot}$. Additionally, the mass spectrum as a whole decays more rapidly than what was seen in GWTC-2 as the new observations in O3b contain more low mass systems.

The evolution of the BBH rate as a function of redshift following equation 3.1 also changes with these new detections. This can be seen in Figure 3.2 along with the equivalent calculation from GWTC-2 in the top panel, and the Madau–Dickinson star formation rate model in the bottom panel. We measure the merger rate the best around $z = 0.2$ as $\mathcal{R}(z = 0.2) = 19 - 42 Gpc^{-3} yr^{-1}$. Additionally, we characterize κ using these latest observations as $\kappa = 2.9^{+1.7}_{-1.8}$ which is a higher values than that assumed from GWTC-2.

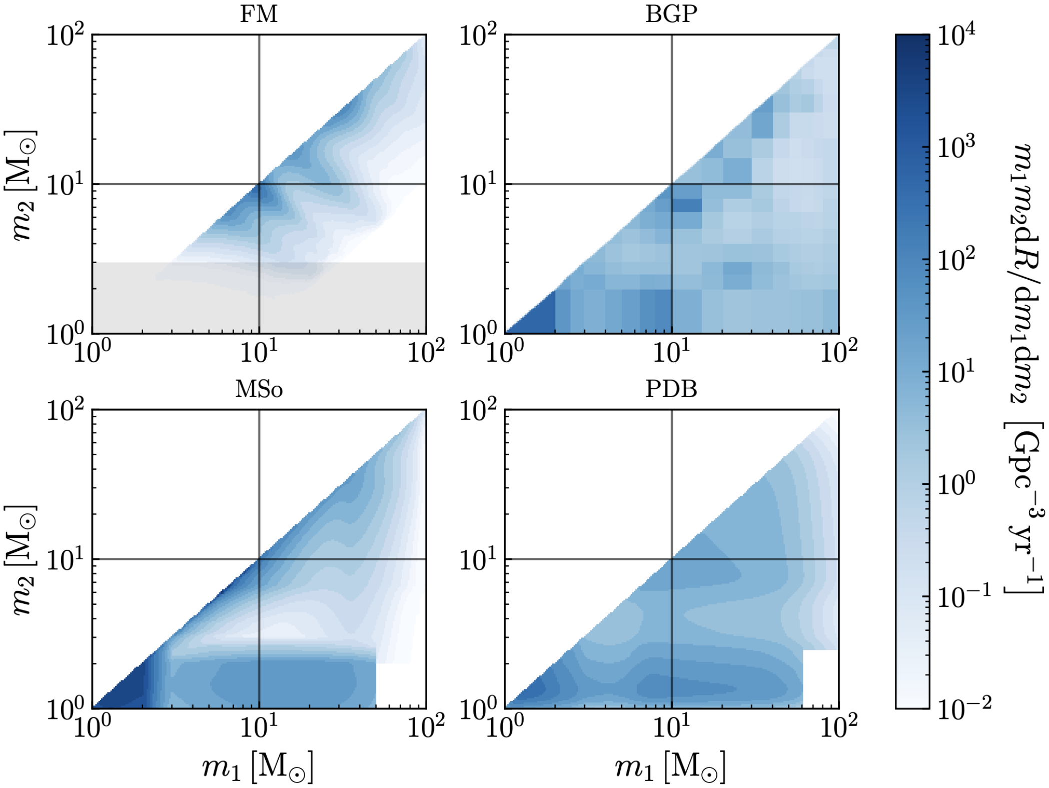


Figure 3.1. Rate density versus component masses for different models inferred from events with FAR < $0.25yr^{-1}$. Top left panel: Rate density computed with the FM model assuming no redshift evolution, for binary black holes only. Top right panel: Rate density inferred with the BGP model using all compact objects. Bottom left panel: Rate density inferred with MS. For mergers involving typical BH, this model strongly favors equal-mass mergers. Bottom right panel: Rate density inferred with PDB. Reprinted from [6]

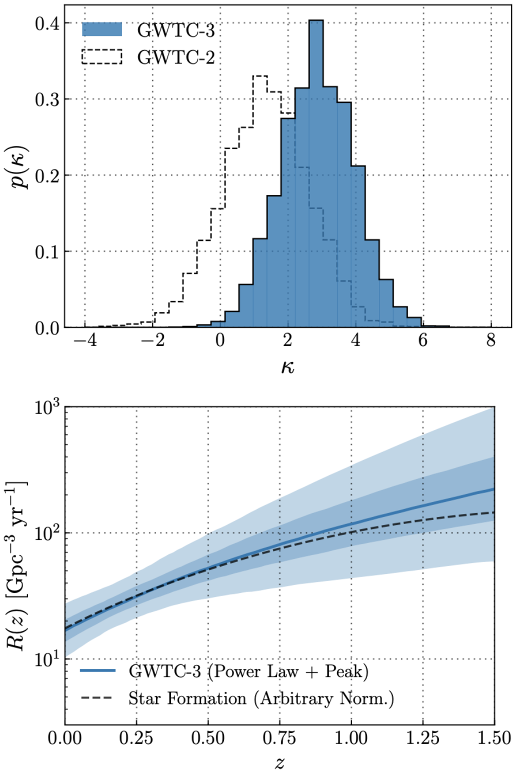


Figure 3.2. Constraints on the BBH merger rate with redshift. Top: Posterior on the power-law index κ governing the BBH rate evolution, which is presumed to take the form given in equation 3.1. The blue histogram shows our latest constraints using GWTC-3 , while the dashed distribution shows our previous constraints under GWTC-2. Bottom: Central 50% (dark blue) and 90% (light blue) credible bounds on the BBH merger rate $\mathcal{R}$(z). The dashed line, for reference, is proportional to the rate of cosmic star formation [7]; we infer that $\mathcal{R}(z)$ remains consistent with evolution tracing star formation. Reprinted from [6]

Chapter 4
Real-Time Gravitational Wave Detection During LIGO, Virgo, and KAGRA's Fourth Observing Run

The LIGO and Virgo collaboration's fourth observing run began on May 24th, 2023 with a planned end date of February 2025. The observing run was split into two halves. The first half, O4a, ran from May 24th, 2023 to January 16th, 2024 with a preceding engineering run from April 26th to May 23rd, 2023. During this time, the LIGO Livingston (LLO) and LIGO Hanford (LHO) detectors operated on average between 140 and 150 Mpc, and the Virgo interferometer did not participate. The second half, O4b, started on April 10th, 2024 and with a preceding engineering run from March 20th to April 10th of the same year. LIGO Hanford, LIGO Livingston, and Virgo plan to participate with LHO and LLO at similar or slightly better sensitivities than O4a, and Virgo operating at around 50Mpc. Unfortunately, KAGRA's plans of joining O4b have been set back due to the January 1st, 2024 earthquake which caused considerable damage in the region, and they now plan to begin operations in 2025.

The same gravitational wave detection pipelines which searched for a wide range of compact objects (AllSky) in O3 are now also participating in O4 including GstLAL [9], PyCBC [56], MBTA [55], SPIIR [57], and cWB [54]. New to O4, however, are low-latency targeted searches for sub-solar mass objects (SSM), operated by GstLAL and MBTA, and early-warning (EW) detection, operated by GstLAL and PyCBC. In addition, PyCBC and MBTA began analyzing single detector candidates for low-mass sources, while GstLAL continues to analyze all single detector data for any source.

Ahead of O4, I worked on the development team for the GstLAL pipeline and for the data quality pipeline iDQ (see Ch 6) implementing bug fixes, general improvements, and more in preparation for the observing run. I additionally lead the effort on operating the pre-engineering analyses discussed in detail in section 4.2.2 before the run began. During O4, I was a lead member of the low-latency operating team for both GstLAL and iDQ, ensuring optimal uptime and accurate scientific output during live data collection. As part of this effort, I spear-headed the implementation of low-latency injection recovery with GstLAL for the first time in O4a, including adding the infrastructure necessary for availability of these injections, as discussed in more detail in Ch 5.

The following chapter summarizes the initial results of the low-latency O4 gravitational wave search using information available to the public at the time of writing. We begin by describing detector upgrades ahead of O4, and the improvements to the GstLAL All-Sky detection pipeline. We then briefly discuss the implementation of a GstLAL analysis which operated before the official start dates of the observing run. Finally, we summarize the performance of the GstLAL All-Sky analysis ahead of the fourth observing run, and discuss initial publicly-available detection results from O4a.

4.1 Detector Upgrades

The LIGO Hanford (LHO) and Livingston (LLO) detectors both experienced hardware upgrades between O3 and the start of O4. These are discussed in detail in [8], but a summary is provided here. Both sites increased the input laser power, and replaced and upgraded a subset of mirrors and mirror coatings. As discussed in section 1.3.2, increasing the input laser power increases sensitivity at high frequencies and decreases it at lower frequencies. To combat this effect, frequency dependent squeezing was implemented at both sites.

Ahead of O3, squeezed light sources were introduced as discussed in Chapter 3. These squeezed light sources reduced the photon shot noise at the expense of increasing radiation pressure noise across the frequency spectrum. Radiation pressure noise dominates at low frequency, so increasing its effects in that regime decreases the overall sensitivity of the detector in that region. A frequency-dependent squeezer mitigates this effect, and was implemented at both LHO and LLO ahead of O4a. Frequency-dependent squeezing reduces the radiation pressure noise at low frequency and increases the shot noise at that frequency. Unlike frequency-independent squeezing, however, the opposite is implemented at higher frequencies resulting in an improvement in sensitivity across the frequency spectrum.

Each of the detector sites also face unique challenges, and therefore additional unique improvements were made at each as well. At Livingston, glitches caused by logging were one of the most frequent noise transients present in the detector as classified by GravitySpy [25]. The ground motion created by logging practices caused resonances in the optics designed to prevent scattered light from recombining back into the main laser beam. Therefore, dampeners were added to these optics in order to decrease the response of these resonances.

At Hanford, meanwhile, electronic noise from changes in the local electric field was mitigated. The system suspending the main mirrors in the interferometer has four stages and the final stage of this system implements electrostatic drives (ESDs) to hold the mirror in place via electric force. However, changes in the local electric field can make the force applied by the ESDs vary, resulting in slight variations in the positions of the main mirrors. The wires attached to the ESD systems were grounded, making then more insensitive to these local variations.

The improvements resulted in an increase in sensitivity at both sites between O3 and O4. The differences between the PSD of the detectors representing this improvement can be seen in Figure 4.1.

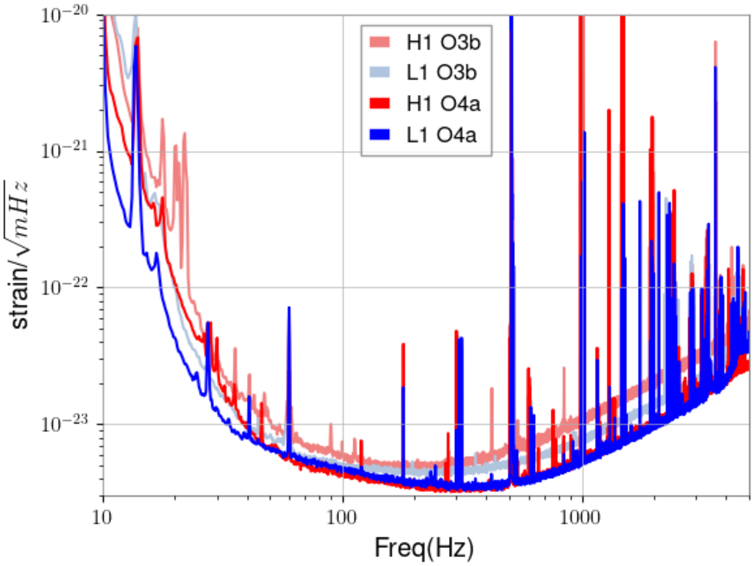


Figure 4.1. Noise sensitivity curves of LIGO-Hanford(H1) and LIGO-Livingston(L1) at the end of the third observing run (O3), and at the start of the fourth (O4a). Reprinted from [8].

4.2 Methodology of the Real-Time GstLAL Analysis ahead of O4

4.2.1 Improvements Ahead of O4

4.2.1.1 Signal Contamination Removal

As discussed in 2.1.5, histograms of single detector events during coincidence time are populated to support the calculation of the $P(\vec{\rho}, \vec{\xi^2}|\theta, n)$ term in the likelihood ratio. However, occasionally only a single detector will register a real gravitational wave event during observing time when more than one detector is collecting science quality data. This can happen if, for example, an event is in one detector's blind spot, or if one of the detectors participating has a higher sensitivity than another. In this case, the analysis by default marks this event as noise, and adds the corresponding samples to the noise histogram. Samples corresponding to a real signal fall into a space in the noise histogram that is sparsely populated, and application of a KDE exacerbates this effect. An example of this can be seen in the left of Figure 4.2.

The noise background histograms track the frequency at which noise events lie in the SNR, ξ^2 space. Therefore, adding true-signals to the noise histogram makes it appear that noise events occur in the same SNR, ξ^2 space as real gravitational waves. If not rectified, this contamination from real events can therefore result in a decrease of the likelihood ratio for true gravitational wave events by falsely inflating the $P(\vec{\rho}, \vec{\xi^2}|\theta, n)$ term.

To mitigate this, a new feature was added to the analysis which allows these samples to be removed by the operating team [49]. Now, any time an event is uploaded to GraceDB, a record

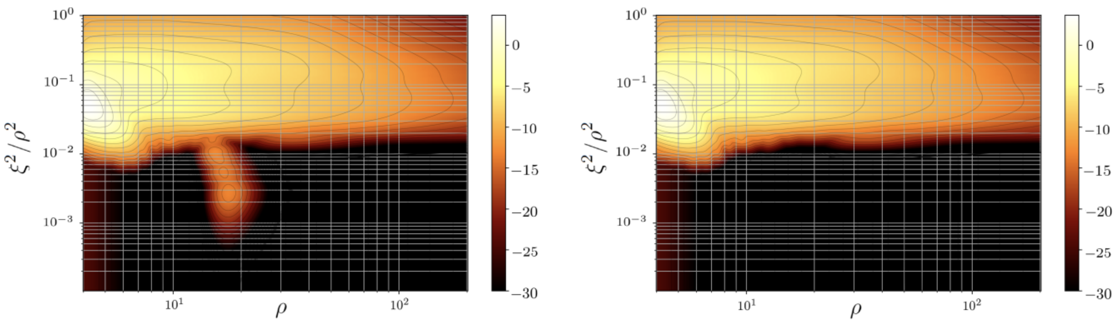


Figure 4.2. Noise background distribution from a single background bin in the GstLAL analysis. Left: The bin when contaminated with a real gravitational wave event, thereby extending the background support outside of its regular space. Right: The same background bin which has had the counts removed for the gravitational wave event in question.

of the samples added to histograms around the event time is kept. Then, if a single-detector candidate is registered with a significance above the open public alert threshold, operators can manually mark these event times for signal contamination removal. Times flagged following this method have samples removed in a ten second window around the event time across template bins thereby removing the effects of real signals from the noise background. The end result of sample removal can be seen on the right of Figure 4.2.

4.2.1.2 Combo-ξ^2 Implementation

Recall the ξ^2 signal consistency check from 2.1.5, denoted as ξ_t^2 here. This was a consistency check in the time domain and compared the SNR timeseries of the incoming signal when matched with a template to the SNR timeseries expected from the template. However, there can also be signal consistency checks in other domains. The ξ_b^2, or bank-ξ^2, check compares the complex SNR timeseries (defined in 2.25) from the signal and template in question with same result from the signal and other templates in the same bin. While ξ_t^2 implemented the residual SNR from the auto-correlation of the template, ξ_b^2 estimates the residual SNR from the match of the template α with its neighbors weighted by the peak of the detected SNR timeseries in template α, or [49]:

$$\xi_b^2 = \sum_{\beta \in \theta} \frac{|\rho_\beta[0] - \rho_\alpha[0](h_\beta|h_\alpha)|^2}{2 - |(h_\beta|h_\alpha)|^2} \tag{4.1}$$

where

$$(h_\beta|h_\alpha) = \sum_\tau h_\beta[\tau]h_\alpha^*[\tau] \tag{4.2}$$

Note that the summation here is over the nearby templates in the bank, and not time as it was for ξ_t^2.

ξ_b^2 can then provide complimentary information to ξ_t^2, and the two were combined ahead of

O4 to make one ξ^2 variable [49]. This combination, called ξ^2_{tb} is given as:

$$\xi^2_{tb} = \frac{1}{N_t + N_b}(N_t\xi^2_t + N_b\xi^2_b) \tag{4.3}$$

where N_b and N_t are the denominators, or normalization factors, of 4.1 and 2.28, respectively. This new ξ^2_{tb} value was tracked over the course of O4a, but was not implemented in the likelihood ratio. There are plans to do a full comparison of the likelihood ratio with and without the implementation of ξ^2_{tb} using this data to determine if it will be used in the future.

4.2.2 Pre-Engineering Analyses

Ahead of an observation run, before even an engineering run begins, detector commissioners perform maintenance, implement features, and test configuration changes. During this pre-engineering time, detectors may be operational, but the quality of the data and the detector's sensitivity is not guaranteed. Pre-engineering data has never historically been utilized by detection pipelines because of the data quality and availability uncertainty associated with it.

However, gravitational waves can arrive at any time and another event like GW170817 could occur during this under-utilized observing time. An exceptional gravitational wave detection during this time would require additional vetting and verification which exceeds that of any detection made during production-quality observing time, but the cost would not outweigh the scientific gains of another multi-messenger event detection. Therefore, ahead of both halves of O4, the GstLAL pipeline dedicated resources to running one half of the full checker-boarded analysis during these pre-engineering times. The considerations for data availability and quality required for analysis of this data as well as preliminary findings are discussed in the following sections.

4.2.2.1 Data Acquisition

Ahead of O4, the architecture of data streaming changed for the low-latency analysis use case. Data is now streamed from the detectors to RAM on computing clusters in gravitational wave frame files (.gwf) files. Each file contains a variety of information about the strain data, detector state, and data quality in one second durations. The GstLAL analysis was updated to read this data as it becomes available and extract the relevant information.

Typically, it is guaranteed that these data files arrive from every detector for every second of the observing run – even when the detector is not producing science quality data. During pre-engineering times, however, the detectors and data distribution system are not guaranteed to be fully functional. Intermittently, then, there can be data distribution issues which cause data from one, or many, detectors to either arrive at some delay, or stop arriving all together.

In these cases, it's desirable that detection pipelines continue to produce scientific output on any remaining available data. Ahead of O4, functionality was added to the GstLAL analysis that adds support for this case. If the analysis does not receive new data for longer than a tunable wait time, then it creates null data in its place. In this way, the pipeline as a whole can continue to produce output with an additional latency equal to the wait time instead of halting production all together. Ahead of O4, this wait time was set to one minute to allow for minor fluctuations in data availability, while maintaining optimal uptime.

4.2.2.2 Data Quality Considerations

Gravitational wave data contains state vectors which provide information about whether or not the data is considered production-level quality. During pre-engineering times, the state vector is never set to indicate that the data is ready for analysis, so the GstLAL pre-engineering analysis must operate ignoring this state vector information. Without this, however, all information about the state of the detector is lost and the data is analyzed no matter the detector state.

Frequently, commissioning activities cause extremely loud and long duration glitches which would typically not be present during this time. The internal gating scheme in GstLAL is designed to handle brief departures from the norm as described in 2.1.3, and it handles the majority of these transients well. However, when these transients are extremely frequent, the frequency spectrum of the data can have large amplitude spikes at the frequency corresponding to the local glitches. Then, when whitened, the average must come to 1, meaning that the power in this one frequency bin is spread into the surrounding bins. This can cause the noise floor of the detector to seem artificially inflated, thereby creating a range much smaller than expected. Additionally, the calibration of the strain data during these times is highly uncertain. The calibration is typically not yet tuned to any changes made to the detector since the last observing run, and commissioning activities cause highly irregular behavior in the output channels used to reconstruct the strain. These combined effects can make the calibration errors large, and therefore introduce systematic error of factor a few in the range calculation (see Eq 1.29) as well.

A skewed range calculation then has a cascading effect into the likelihood ratio and significance estimations. Therefore, an additional gate was added specifically for use during this pre-engineering time which removes data based on the calculated range, effectively acting as a state vector for the analysis. A minimum range of 10Mpc was enforced with a maximum of 600Mpc. For reference, the standard range of the detectors during O4 was between 100 to 150Mpc, so this was a fairly conservative implementation mainly used to bound the most extreme calibration uncertainties.

Maintenance activities and atypical glitches can also pollute the noise background histograms in a similar method as real gravitational wave signals. The noise triggers generated during these times will not accurately represent the typical noise parameter space, and can therefore skew the likelihood ratio and FAR calculations for any real gravitational wave. Ahead of O4b, however, this effect was mitigated by initially populating the background histograms with data collected over the course of O4a. This allowed the bulk of the histogram to be representative of typical detector behavior, and then adjusted by the most recent data over time. Therefore, the likelihood ratio and FAR estimations were more robust for the duration of the analysis.

4.2.2.3 Pre-Engineering Results

During the pre-engineering analysis ahead of O4a, three possible candidates were identified. Two of these generated a significant SNR in one of the two detectors operating at the time, and without further follow-up, this increases the chance that these are false alarms. The third candidate, however, was coincident between the detectors on April 24th, 2023 with a FAR that crossed the threshold for typical inclusion as a gravitational wave candidate. A spectrogram of this event similarly shows the chirp morphology of a GW in both detectors at the event time. Unfortunately, this pre-engineering analysis was not configured to upload candidate events to GraceDB for analysis, and so no notification of this candidate was distributed to the wider community. As none of these events were thought to contain exceptional source properties, there was no additional follow-up by the collaboration, and none is expected. During the engineering run itself, the pre-engineering analysis also identified two additional candidates, along with other search pipelines running at the time. Both candidates passed the typical FAR threshold, and recovering events in coincidence with other searches validates the effectiveness of the pre-engineering search. At the time of writing, details of these pre-engineering events are unfortunately not publicly available, and therefore not discussed in more detail here.

The pre-engineering analysis ahead of O4b is ongoing at the time of writing. It is currently configured to upload findings to GraceDB, and has been bootstrapped with the noise background from O4a, making it more likely to accurately recover events. Before the engineering run began, it recovered 5 significant candidates and initial follow-up of these events is promising.

4.3 O4a results

4.3.1 Performance of GstLAL Ahead of O4a

As discussed in the previous sections, the low-latency GstLAL All-Sky search pipeline received several updates ahead of O4a. We will summarize here the performance of those updates on a known set of data following [9]. In the preparation for O4, a Mock Data Challenge (MDC) was designed to replay data from O3 as if it were live from the detectors. The original strain data was available alongside simulated waveforms with known parameters embedded in strain data, or an injection set. This allowed searches like GstLAL to operate their low-latency configurations over this data as if it were real-time thereby testing the sensitivity and performance of their pipelines.

Here, we test the GstLAL analysis using an MDC covering forty days originally recorded from Jan 5th, 2020 to Feb 14th, 2020 and replayed in April of 2023. This data contains strain information from LHO, LLO, and Virgo, and we report on operation by GstLAL analyzing strain data from all three detectors, despite O4a only have LHO and LLO participation. The details of the injection set used in this MDC can be found in [65], but it generally mimics the expected astrophysical distribution. This results in a total of 5,000 injections throughout the duration which lie in the parameter space covered by the template bank (see section 2.1.1) and beyond it.

The sensitivity of searches is typically measured in terms of the sensitive time-volume, or ⟨VT⟩. In order calculate ⟨VT⟩, we first define the efficiency, ϵ of a search. The efficiency is calculated per FAR threshold, and per luminosity distance bin, D_L, as [9]:

$$\epsilon(\mathrm{FAR}, D_L) = \frac{N_{\mathrm{found}}(\mathrm{FAR}, D_L)}{N_{\mathrm{total}}} \tag{4.4}$$

where $N_{\mathrm{found}}(FAR, D_L)$ is the number of recovered injections at the FAR and D_L bin, N_{total} is the total number of injections irrespective of FAR and distance, and the D_L bins are created

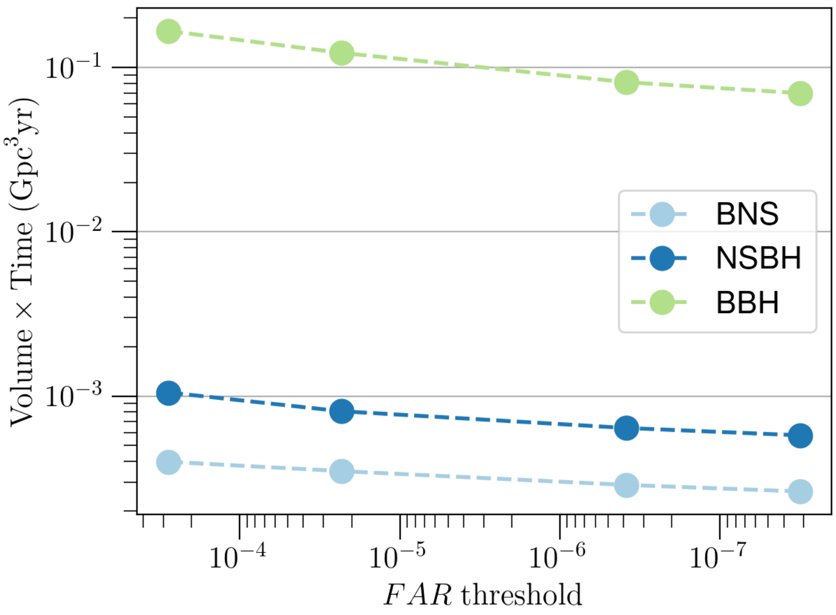


Figure 4.3. ⟨VT⟩ for the GstLAL analysis over the course of an MDC in each source class for a variety of FAR thresholds. Note that the drop in ⟨VT⟩ in the BNS and NSBH regions are not due to their efficiency, but their source classes. Reprinted from [9]

by dividing the total injection set by the luminosity distance of the injected waveforms. We additionally enforce that the injections counted in N_{total} and N_{found} have an injected decisive SNR of at least 8 where the decisive SNR is given by the SNR of the second most sensitive detector participating. This ensures that injections which we would not expect to recover, such as distance BNS, are not included in this estimate. The ⟨VT⟩ is then given as:

$$\langle \text{VT(FAR)} \rangle = 4\pi T \int_{D_{low}}^{D_{high}} \epsilon(\text{FAR}, D_L) D_L^2 dD_L \tag{4.5}$$

where D_{low} and D_{high} are the nearest, and farthest luminosity distances in the injection set, and T is the observing time. In the limiting case of $\epsilon(\text{FAR}, D_L) = 1$, then this returns the expected ⟨VT⟩ of the injection set, $\langle\text{VT}\rangle_{inj}$. The ⟨VT⟩ of the analysis is then split into multiple source classes by only considering injections in the mass range given by the source classification. The result of this over the course of the entire MDC can be seen in Figure 4.3.

This figure shows that the ⟨VT⟩ decreases with the increasing FAR threshold. Fewer injections will be found at increasing confidence of the increased FAR threshold, and therefore the fraction of found injections will decrease along with the ⟨VT⟩. A decreasing ⟨VT⟩ with decreasing source class mass is also clear in the figure. Heavier mass systems at similar distances radiate more energy in gravitational waves, and therefore induce a larger strain in the detectors and a generally larger SNR in searches. A heavier mass system then can be detected at a larger distance with the same SNR as a lower mass system at a smaller distance. This results in the $\langle\text{VT}\rangle_{inj}$ for lower mass systems being smaller at the outset, so even if the efficiency in

FAR(Hz)	BNS	NSBH	BBH
2.78e-4	0.95	0.77	0.87
2.31e-5	0.94	0.71	0.84
3.85e-7	0.89	0.65	0.75
3/16e-8	0.86	0.62	0.69

Table 4.1. Injection efficiency of several source classes by the GstLAL analysis at a variety of FAR thresholds. Re-created from [9]

that mass range is higher, the $\langle VT \rangle$ will still be worse. Such was the case here, as can be seen in Table 4.1 which shows the efficiency of these same source classes.

Table 4.1 shows that the efficiency of the recovered injections decreased with increasing FAR, again because of the increasing confidence. It also shows that our analysis had the highest efficiency for BNS systems, and the worst efficiency for NSBH systems. We found over the course of the MDC that there were several high SNR injections which we would expect to recover that were actually missed by the analysis. These typically lived in the high mass BBH and NSBH space, and therefore negatively affect the efficiency of the NSBH and BBH regions.

Follow-up showed that these injections fell outside of the SNR-ξ^2 region used in the signal model of the likelihood ratio. As described in 2.1.5, the ξ^2 test depends on a mismatch between a template from the bank and the recovered SNR timeseries, but the template bank is sparse in the high mass region. Therefore, the allowed mismatch had evidently been set to a range small enough that it was excluding these events. As a result, this mismatch range was tuned ahead O4a to allow for more dissimilarities in these regions while taking care not to make it so wide as to allow noise to fall within it.

There are additionally nine known gravitational wave events previously published in GWTC-3 [13] which occur during the time period covered by this MDC. Details of the recovery of these events are given in [9], but we will discuss a summary here. Three of these events were detected by the original GstLAL online analysis operating during O3 at a FAR threshold of less than one per year. Two others were detected by the O3 online analysis at a FAR too high for significance, but low enough to be uploaded to GraceDB.

Seven events, including all those detected by the O3 online analysis, were recovered by the MDC analysis ahead of O4 at or above the significant FAR threshold showing a performance at par or better than the O3 online analysis. Of the two remaining events, both were recovered by the MDC analysis, but not significantly. However, the fact that either were recovered, even at low significance, is an improvement over the O3 online search.

4.3.2 Public O4a Results

There were 81 total low-latency public events during O4a which were uploaded to GraceDB [14], and which all had subsequent Gamma-ray Coordinate Network (GCN) notices sent to the public. If all pass the candidacy thresholds, these nearly double the total number of gravitational wave candidates to date. There were additionally 11 retractions over the course of O4a, as displayed in Table 4.2. A retraction is an event for which there was a public alert

Superevent ID	Pipeline	P(astro)	FAR(Hz)	Time to Retraction (Hours)	GCN
S231112ag	GstLAL Allsky	1.0 BBH	1.061e-14	2	[66]
S231030av	SPIIR EW	0.93 BNS	4.215e-08	0.33	[67]
S230918aq	PyCBC EW	0.79 BNS	5.418e-08	0.66	[68]
S230830b	SPIIR Allsky	0.8 NSBH	1.148e-10	1	[69]
S230810af	SPIIR EW	0.99 BNS	2.905e-08	0.66	[70]
S230808i	cWB BBH	0.99 Terr	6.851e-11	4	[71]
S230715bw	SPIIR Allsky	0.91 NSBH	7.843e-09	11	[72]
S230712a	GstLAL Allsky	0.99 BBH	3.269e-15	1	[73]
S230708bi	GstLAL Allsky	0.99 BBH	1.113e-09	0.66	[74]
S230622ba	MBTA	0.87 BBH	5.180e-08	0.5	[75]
S230524x	PyCBC EW	0.75 BNS	7.224e-08	10	[76]

Table 4.2. Retracted public alerts over the course of O4a labeled by their GraceDB [14] superevent ID.

sent, but which was later confirmed to be of terrestrial origin. GstLAL contributed 3 of these, while SPIIR EW, SPIIR AllSky, and PyCBC EW all contributed 2, and cWB and MBTA just one. Of these retractions, only one was labeled confidently as a terrestrial event via its low-latency p_{astro} and five of the eleven were labeled as potential binaries containing a neutron star. Gravitational waves from binaries containing neutron stars are the sources most likely to produce EM radiation and are therefore of particular interest to astronomers and external laboratories. It's especially important then that event alerts for these binaries which are found to be of terrestrial origin are quickly retracted by the LIGO collaboration when possible. From Table 4.2, we see that this is generally the case and that most retracted events were retracted within an hour of their original notice. This table highlights the importance of prompt vetting by experts for any low-latency event.

Of the events which were not retracted, there were 81 above the public alert threshold, and these results can be seen in Table 4.3. The participating pipelines each recovered a different number of events at or below this threshold. GstLAL, PyCBC, cWB, MBTA, and SPIIR recovered 62, 38, 36, 19, and 19 events respectively below a FAR threshold of 2 per year and a trials factor of 5 where this threshold is chosen following the latest GW catalogue, GWTC-3 [13]. Above the same FAR threshold, GstLAL, PyCBC, cWB, MBTA, and SPIIR recovered 19, 19, 21, 43, and 34 respectively, and these events are marked in italics in Table 4.3. Note that FAR threshold for public alerts is 2/day and so there will be some events reported in Table 4.3 which no pipeline recovered above the 2 per year threshold.

The p_{astro} value displayed in Table 4.3 corresponds to the with the highest SNR out of all the participating pipeline's triggers for each superevent. All but two of the public events were confidently classified with a p_{astro} consistent with a BBH. S230529ay, aka GW230529, will be discussed in detail in the following section. S230627c, meanwhile, has a p_{astro} of 49% NSBH and 48% BBH reported by PyCBC, but a greater than 99% chance of being a BBH as reported by GstLAL. More follow-up will need to be done on this event to discover its true nature, but the differing reports between pipelines decreases the certainty of a NSBH origin.

Superevent	Instruments	p_{astro}	FAR (Hz)					GCN
			GstLAL	PyCBC	SPIIR	MBTA	CWB	
S240109a	H1	0.995 BBH	7.35e-09					[77]
S240107b	H1,L1	0.966 BBH	*5.83e-08*	*4.76e-06*	*2.16e-06*	*1.28e-06*	*7.34e-08*	[78]
S240104bl	H1	1.0 BBH	3.55e-17					[79]
S231231ag	H1	1.0 BBH	8.35e-15					[80]
S231226av	H1,L1	1.0 BBH	1.11e-50	3.17e-10		1.58e-09	6.65e-11	[81]
S231224e	H1,L1	0.996 BBH	7.46e-26	3.17e-10		1.53e-09		[82]
S231223j	H1,L1	0.999 BBH	1.11e-09	3.01e-08		*6.15e-08*	4.31e-09	[83]
S231213ap	H1,L1	1.0 BBH	6.08e-12	6.34e-10	1.27e-08	*9.73e-08*	3.26e-09	[84]
S231206cc	H1,L1	1.0 BBH	1.93e-35	3.17e-10	5.09e-16	1.58e-09	6.35e-11	[85]
S231206ca	H1,L1	1.0 BBH	7.49e-21	3.17e-10	1.29e-10	2.65e-08	4.45e-10	[86]
S231129ac	H1,L1	0.986 BBH	1.76e-08		*7.64e-05*		*3.5e-06*	[87]
S231127cg	H1,L1	0.996 BBH	5.81e-09	*1.83e-07*	*2.03e-06*	*8.82e-07*	4.4e-09	[88]
S231123cg	H1,L1	1.0 BBH	*0.000186*	3.17e-10			6.48e-11	[89]
S231119u	H1,L1	0.955 BBH	*7.43e-08*		*7.7e-05*	*0.000103*	*2.12e-06*	[90]
S231118an	H1,L1	0.743 BBH	1.71e-12	3.17e-10	*7.48e-08*	1.56e-09		[91]
S231118ab	H1,L1	0.985 BBH	2.82e-10	*9.44e-07*	1.9e-08	*1.07e-07*	1.67e-09	[92]
S231118d	H1,L1	1.0 BBH	1.33e-12	3.17e-10	*3.23e-08*		*1.27e-05*	[93]
S231114n	H1,L1	1.0 BBH	7.06e-12	3.17e-10	3.1e-09	9.38e-09		[94]
S231113bw	H1,L1	0.789 BBH	7.17e-11	1.36e-08	*9.1e-07*	1.57e-09	*4e-05*	[95]
S231113bb	H1,L1	0.965 BBH	*5.6e-08*			*3.21e-06*		[96]
S231110g	H1,L1	0.968 BBH	2.58e-14	3.17e-10	1.93e-08	1.5e-09		[97]
S231108u	H1,L1	1.0 BBH	7.03e-25	3.17e-10	*8.1e-08*	1.34e-09	6.56e-11	[98]
S231104ac	H1,L1	0.996 BBH	5.8e-18	3.17e-10				[99]
S231102w	H1,L1	1.0 BBH	1.68e-15	3.17e-10	5.84e-23	1.68e-09	6.47e-11	[100]
S231029y	L1	1.0 BBH	2.16e-10					[101]
S231028bg	H1,L1	1.0 BBH	7.63e-31	3.17e-10	1.7e-11		6.42e-11	[102]
S231020bw	H1,L1	1.0 BBH	3.45e-10					[103]
S231020ba	H1,L1	0.912 BBH	3.31e-17	1.27e-09	6.38e-13	2.84e-08		[104]
S231014r	H1,L1	0.992 BBH	1.03e-08	*5.2e-07*		*4.08e-06*	*1.98e-06*	[105]
S231008ap	H1,L1	0.999 BBH	1.53e-09	*4.56e-07*	*2.8e-07*	*1.01e-06*	*9.54e-05*	[106]
S231005ah	H1,L1	0.998 BBH	2.05e-09	*2.18e-05*	*7.75e-05*	*2.79e-06*	*5.24e-07*	[107]
S231005j	H1,L1	0.978 BBH	*3.22e-08*	*1.39e-05*		*2.79e-06*	5.7e-09	[108]
S231001aq	H1,L1	0.996 BBH	4.97e-09	6.34e-10	*7.49e-06*	*4.2e-07*	2.6e-10	[109]

(To be continued)

Superevent	Instruments	p_{astro}	FAR (Hz)					GCN
			GstLAL	PyCBC	SPIIR	MBTA	CWB	
S230930al	H1,L1	0.994 BBH	7.38e-09	*0.000145*		*1.51e-06*	*7.13e-07*	[110]
S230928cb	H1,L1	1.0 BBH	1.24e-12	9.5e-10	*2.99e-07*	*1.71e-05*	3.27e-10	[111]
S230927be	H1,L1	1.0 BBH	4.56e-43	3.17e-10	1.65e-24	1.35e-09	6.64e-11	[112]
S230927l	H1,L1	0.976 BBH	2.8e-14	3.17e-10	1.08e-08	*5.41e-08*	1.31e-10	[113]
S230924an	H1,L1	1.0 BBH	1.21e-21	3.17e-10	3.82e-09	1.49e-09	6.59e-11	[114]
S230922q	H1,L1	1.0 BBH	7.72e-13	3.61e-10		*3.97e-05*	1.98e-10	[115]
S230922g	H1,L1	1.0 BBH	1.93e-24			1.91e-08	1.98e-10	[116]
S230920al	H1,L1	1.0 BBH	5.19e-13	3.17e-10	*2.49e-05*	1.42e-08	6.69e-11	[117]
S230919bj	H1,L1	1.0 BBH	5.67e-42	3.17e-10	4.72e-09	1.49e-09	6.69e-11	[118]
S230914ak	H1,L1	0.992 BBH	1.43e-20	3.17e-10	9e-10		6.5e-11	[119]
S230911ae	H1	1.0 BBH	1.89e-12					[120]
S230904n	H1,L1	0.991 BBH	6.16e-13	3.17e-10	*2.64e-06*	2.25e-09		[121]
S230831e	H1,L1	0.985 BBH	1.98e-08	*1.66e-05*	*5.84e-05*	*7.93e-07*	*9.45e-06*	[122]
S230825k	H1,L1	0.998 BBH	2.39e-09			*2.67e-06*	*1.7e-07*	[123]
S230824r	H1,L1	1.0 BBH	6.17e-12	3.17e-10	1.64e-11		6.18e-11	[124]
S230822bm	H1,L1	0.981 BBH	2.58e-08				*1.09e-06*	[125]
S230820bq	H1,L1	0.958 BBH	2.41e-08	*2.1e-06*	*4.23e-08*	*6.06e-06*	9.01e-09	[126]
S230819ax	H1,L1	0.993 BBH	8.84e-09		*4.57e-06*	*2.76e-06*	5.58e-09	[127]
S230814ah	L1	1.0 BBH	1.85e-21					[128]
S230814r	H1,L1	0.932 BBH	7.63e-10	3.17e-09	*4.81e-08*		9.31e-11	[129]
S230811n	H1,L1	1.0 BBH	2.13e-25	3.17e-10	4.63e-21	1.39e-09	6.81e-11	[130]
S230807f	H1,L1	0.953 BBH	*7.14e-08*	*5.56e-05*	*5.75e-05*	*3.6e-05*	*3.06e-06*	[131]
S230806ak	H1,L1	0.997 BBH	2.96e-09	*1.1e-06*	*3.82e-05*	*4.89e-06*	1.41e-08	[132]
S230805x	H1,L1	1.0 BBH	2.72e-10	9.19e-09	*3.08e-05*	3.1e-08	*2.81e-06*	[133]
S230802aq	H1	0.903 BBH	2.23e-08					[134]
S230731an	H1,L1	0.814 BBH	4.57e-27	3.17e-10	3.91e-12	1.43e-09	*4.81e-05*	[135]
S230729z	H1,L1	0.997 BBH	3.39e-09	*8.56e-07*		*1.94e-05*		[136]
S230726a	L1	1.0 BBH	3.83e-14					[137]
S230723ac	H1,L1	0.867 BBH	5.61e-10		*4.57e-05*	*5.33e-08*		[138]
S230709bi	H1,L1	0.997 BBH	3.06e-09	*1.85e-06*	*2.35e-05*	*6.47e-07*	4.93e-09	[139]
S230708cf	H1,L1	0.989 BBH	1.55e-08		*6.63e-06*	*3.48e-06*	*2.11e-05*	[140]
S230708z	H1,L1	0.954 BBH	*7.03e-08*			*3.37e-06*	*4.53e-08*	[141]
S230708t	H1,L1	0.973 BBH	*4.33e-08*			*6.19e-07*		[142]

(To be continued)

Superevent	Instruments	p_{astro}	FAR (Hz)					GCN
			GstLAL	PyCBC	SPIIR	MBTA	CWB	
S230707ai	H1,L1	0.951 BBH	1.53e-08	*4.59e-08*	*6.49e-06*	1.38e-08	3.84e-10	[143]
S230706ah	H1,L1	0.973 BBH	*4.26e-08*	*2.5e-06*	*0.000101*	*2.08e-06*		[144]
S230704f	H1,L1	0.997 BBH	2.82e-09			*8.56e-08*	*9.37e-06*	[145]
S230702an	H1,L1	1.0 BBH	1.53e-12	7.92e-09	*1.74e-06*	*7.61e-07*	5.65e-09	[146]
S230630bq	H1,L1	0.968 BBH	2.66e-09	1.55e-08	*3.39e-06*	7.73e-09		[147]
S230630am	H1,L1	0.983 BBH	2.41e-08			*6.78e-05*	*1.62e-07*	[148]
S230628ax	H1,L1	1.0 BBH	6.6e-32	3.17e-10	1.82e-13	1.39e-09	5.72e-10	[149]
S230627c	H1,L1	0.492 NSBH	9.59e-45	3.17e-10	1.23e-28	*8.64e-08*	3.98e-10	[150]
S230624av	H1,L1	0.953 BBH	2.96e-11		*6.16e-07*	1.3e-08	3.98e-10	[151]
S230609u	H1,L1	0.961 BBH	2.98e-15	6.97e-09	*2.96e-07*	1e-08	1.35e-09	[152]
S230608as	H1,L1	1.0 BBH	1.37e-10		*3.36e-06*	*1.34e-07*	8.58e-10	[153]
S230606d	H1,L1	0.999 BBH	*5.53e-08*	1.14e-08	*4.4e-07*	1.34e-08	1.75e-08	[154]
S230605o	H1,L1	0.988 BBH	1.71e-12	6.34e-10	2.67e-09	4.52e-09		[155]
S230601bf	H1,L1	1.0 BBH	5.07e-11	3.17e-10	1.71e-15	*9.8e-08*	3.93e-10	[156]
S230529ay	L1	0.624 NSBH	3e-08	1.98e-10				[157]

Table 4.3: Gravitational-wave triggers which received public alerts in low-latency over the course of O4a from all participating pipelines. Italics denote where a pipeline uploaded an event to GraceDB [14], but it did not pass the 2/year FAR threshold.

Additionally, of these 81 events, 17 of them (or 21%) were identified only by GstLAL above the FAR threshold of 2 per year, and 8 of these were found only by GstLAL. These GstLAL events are shown again in Table 4.4 for ease of the reader. Of the 8 events which only had GstLAL uploads, 7 of them were single detector events. In O3, GstLAL was the only pipeline which searched for these events in low-latency, but in O4, PyCBC and MBTA also began analyzing events during single detector time for systems under $5M_{\odot}$. The only event discovered during single detector time in that mass range was the NSBH S230529ay, found by GstLAL and PyCBC, and it is discussed in further detail in the following section. Otherwise, GstLAL continues to be the only pipeline searching for single detector events at high mass, and we report 7 events in this mass range all with above a 90% probability of being from a binary black hole source.

Of the other 10 events which were identified above threshold by GstLAL and below threshold by others, all were detected during coincident time between LHO and LLO. Differences in the sensitivities of the pipelines and how they estimate FAR may account for the differences in significance assigned by pipelines for these events, as was the case in O3.

Superevent ID	Instruments	p_{astro}	FAR(Hz)	Sub-threshold events?
S240109a	H1	0.995 BBH	7.35e-09	No
S240104bl	H1	1.0 BBH	3.55e-17	No
S231231ag	H1	1.0 BBH	8.35e-15	No
S231029y	L1	1.0 BBH	2.16e-10	No
S231020bw	H1,L1	1.0 BBH	3.45e-10	No
S231014r	H1,L1	0.992 BBH	1.03e-08	Yes
S231008ap	H1,L1	0.999 BBH	1.53e-09	Yes
S231005ah	H1,L1	0.998 BBH	2.05e-09	Yes
S230930al	H1,L1	0.994 BBH	7.38e-09	Yes
S230911ae	H1	1.0 BBH	1.89e-12	No
S230825k	H1,L1	0.998 BBH	2.39e-09	Yes
S230814ah	L1	1.0 BBH	1.85e-21	No
S230806ak	H1,L1	0.997 BBH	2.96e-09	Yes
S230729z	H1,L1	0.997 BBH	3.39e-09	Yes
S230726a	L1	1.0 BBH	3.83e-14	No
S230723ac	H1,L1	0.867 BBH	5.61e-10	Yes
S230704f	H1,L1	0.997 BBH	2.82e-09	Yes

Table 4.4. GW candidates identified in low-latency by only GstLAL, or by only GstLAL above a FAR threshold of 2 per year.

4.3.2.1 O4a NSBH Events: GW230518 & GW230529

There have been two events over the course of O4a which have a p(astro) with significant contribution in the BNS or NSBH class: GW230518 & GW230529. GW230518 was detected in low-latency by GstLAL, MBTA, and PyCBC. The FAR from GstLAL and PyCBC were 1.956e-22 and 3.218e-10, respectively, well above the threshold of public alerts and inclusion in previous GW catalogues. Additionally, the p_{astro} of this event was reported as 86% probability of being an NSBH, and 10% of Terrestrial with a 100% probability of having a NS component if real. Unfortunately, this event occurred during the engineering run ahead of O4a, and so has been delayed while a full analysis of the data quality is complete. Initial tests show that despite being during an engineering run, the data quality during this time is good and so we will consider it here as a possible detection.

Of the 81 candidates during the official observing run, GW230529 is the only one reported as having a p_{astro} with significant contribution in the BNS or NSBH class. This event was the first to cross the public alert threshold during the official observing run, and was a single

Primary mass $m_1/M_\odot$	$3.6^{+0.8}_{-1.2}$
Secondary mass $m_1/M_\odot$	$1.4^{+0.6}_{-0.2}$
Mass ratio $q = m_2/m_1$	$0.39^{+0.41}_{-0.12}$
Total mass $M/M_\odot$	$5.1^{+0.6}_{-0.6}$
Chirp mass $\mathcal{M}/M_\odot$	$1.94^{+0.04}_{-0.04}$
Detector-frame chirp mass $(1 + z)M/M_\odot$	$2.026^{+0.002}_{-0.002}$
Primary spin magnitude χ_1	$0.44^{+0.40}_{-0.37}$
Effective inspiral-spin parameter χ_{eff}	$-0.10^{+0.12}_{-0.17}$
Effective precessing-spin parameter χ_p	$0.40^{+0.39}_{-0.30}$
Luminosity distance D_L/Mpc	201^{+102}_{-96}
Source redshift z	$0.04^{+0.02}_{-0.02}$

Table 4.5. Source properties of GW230529, as quoted in [15].

detector event in Livingston, again enforcing the importance of analyzing single detector candidates. A full publication dedicated to this event's analysis and significance is given in [15], and we highlight key results here.

The inferred source properties of GW230529 are shown in Table 4.5. This candidate is an asymmetric event with primary masses solidly in the $< 5M_\odot$ regime. The primary mass is consistent with a black hole living in the low mass gap at a 99% certainty that it is less than five solar masses while the secondary mass is consistent with a neutron star. The second mass has a peak around $1.4M_\odot$, but has a large upper confidence limit leading up to $2M_\odot$.

Additionally, an analysis of the tidal effects which would be present in a GW signal from a neutron star are unfortunately inconclusive. As will be discussed in detail in Chapter 8, the tidal deformability of a neutron star tells us how much its shape will be deformed in tidal fields. If the tidal deformability is high, this corresponds to stars that deform easily, while compact objects like black holes have a tidal deformability of zero. The tidal deformability comes in at high PN order, and is a very sub-dominant effect, so extraordinary sensitivity is required to measure its affect on a waveform to any accuracy. Unfortunately, the sensitivity of the detectors limits our ability to constrain the tidal deformability in this case, and its tidal deformability is completely unconstrained – that is, its distribution is uninformative [15].

Treating this event as its own class of CBCs, we can also calculate the rate of mergers like it by considering it a single Poisson distributed event over the course of O1 to the first two weeks of O4a. This gives a rate of $\mathcal{R}_{230529} = 55^{+127}_{-47} Gpc^{-3} yr^{-1}$ [15]. We can then use the same treatment, but treat this event with the two NSBH events in O3, GW200105 and GW200115, as three Poisson distributed events over the same duration. This gives a rate of NSBH to this point as $\mathcal{R}_{\rm NSBH} = 85^{+116}_{-57} Gpc^{-3} yr^{-1}$ [15]. Additionally, we can also include a population of sub-threshold GstLAL triggers which fall in the NSBH mass parameter space for a full accounting of all potential NSBH up to this point. This gives a slightly higher rate of $\mathcal{R}_{\rm NSBH} = 94^{+109}_{-64}$ [15].

We can further examine the effects of this event on the NSBH and CBC population properties. In the first analysis, we consider the NSBH-POP model [158], a parametric model designed to specifically constrain the mass and spins of the NSBH distribution assuming that all binaries considered have a black hole primary and neutron star secondary mass. With this model, the inferred minimum masses of black holes in NSBH systems decreases when including GW230529. GW230529 contains the lightest BH component of NSBH events recorded to date, so a decrease on the minimum BH components in NSBH systems is a natural conclusion with the inclusion of this event. Additionally, of particular interest is the mass range between 3 and 5 which represents the mass gap between neutron star and black hole populations. If we instead implement mass models which represent the full CBC space without assuming classifications, the merger rates of CBCs with one component in this mass gap range is increased with inclusion of GW230529. This, again, is a natural result as there have been very few detections with masses in this range, the other most notable event being GW190814 [159]. However, GW230529 is mostly consistent with the population from previous CBC candidates while GW190814 was an outlier for BBHs because of its small secondary mass. That said, GW190814 was one of the first events to suggest that the mass gap space is populated with black holes, and GW230529 strengthens this claim.

Chapter 5
Towards Exclusive Low-Latency Analysis of Gravitational Waves

5.1 Introduction

As discussed in Chapter 1, the low-latency analysis of gravitational waves is a crucial part of the multi-messenger astronomy ecosystem. Prompt notification on any gravitational wave event involving a neutron star is key to capturing the full electro-magnetic evolution of such an event by astronomers. To this end, the LIGO collaboration dedicates years of person-power and millions of computing hours over the course of each observing run operating these real-time gravitational wave detection pipelines. After an observing run ends, the data is calibrated, cleaned and then re-filtered with offline detection pipelines in an effort to re-analyze the entire set with improved sensitivity. This effort is effective, as demonstrated by the 17 additional gravitational wave candidates identified offline in O3, but it incurs further computational and person-power costs.

However, if data were originally filtered in low-latency with minimal loss of both sensitivity and uptime, offline re-filtering of the data would be redundant and this additional cost could be mitigated. To this end, we propose two changes to the GstLAL online analysis infrastructure as well as a new method for re-assigning the significance of online candidates without any re-filtering of the data. We then investigate the performance of these methods compared to a full offline treatment and show that initial results prove promising.

5.2 Low-Latency Injection Availability

Offline detection pipelines analyze a simulated injection set embedded in real strain data to compare the performance of the pipeline to what is expected based on the properties of the injection set. Historically, low-latency detection pipelines were not able to do the same because the infrastructure to support it was not available. Some pipelines, such as MBTA [55], have developed strategies to read injection data off of disk, and insert them into their analysis in real-time. While functional for a single pipeline, there are many detection pipelines simultaneously analyzing data in low-latency during a given observing run. Each pipeline could implement their own methods and injections such as MBTA, but the results would be difficult to compare across pipelines as uniformity would not be guaranteed. However, a process upstream of the detection pipelines which made and distributed injection files to all pipelines would mitigate these issues.

The injection streamer proposed here functions as that upstream process. This software requires two inputs: low-latency strain data, and injection waveform data. With this method, the injection data can be made well ahead of an observing run, and timestamped to cover the entire observation period. For example, in preparation for O4b, these injection frames for the entire observing run were made months in advance to reduce workload during low-latency operations. The low-latency strain data required is the same as that typically analyzed by detection pipelines and is available in one second files distributed to a computing center from the detector sites. The injection streamer then adds this real-time strain data to the appropriately timestamped injection from disk, and outputs one second files containing the injections embedded in the real-time data. These output files are formatted identically to a typical strain data, and contain the original strain data, the new injection data, and copies of the original state vectors, and data quality information. The entire process takes a second, so the injection files are available at a one second delay compared to the strain. The output files are then distributed like the original strain data to remote computing nodes across the LIGO Collaboration via the IGWN Low-latency Data Distribution Common.

Any injection set can be distributed in this manner which opens up avenues of low-latency analysis not previously available. Not only can detection pipelines assess the performance of their pipelines in real time, but full injection sets necessary for rates and population estimates could also be analyzed in this manner. Additionally, injection sets targeting particular sources such as sub-solar mass or intermediate black hole sources could be made available separately, thereby supporting a wide variety of detection pipelines at once during an observation run.

5.3 Methodology

5.3.1 Re-ranking Procedure

As described throughly in Chapter 2, the GstLAL analysis implements a likelihood-ratio to assess the significance of gravitational-wave candidate events based partially on the background noise information collected. In the low-latency online operating mode, each candidate event is only ranked based on the background information collected up to that point by the analysis. In the high-latency offline mode, each gravitational wave candidate is ranked with the background noise information from the detectors during the entire observing time. Previously, in order to get a more accurate ranking of candidates, an offline analysis had to be performed, incurring a large computational cost to re-filter all of the available data. In this work, and in the fourth observing run, we instead apply a re-ranking procedure to re-assess the significance of the candidates produced by the online analysis.

The online analysis filters data based on a template bank of waveforms divided into background bins. When the SNR output of a background bin crosses a threshold of 4, a triggered is produced. The result of filtering in low-latency then is a collection of triggers created by each background bin. The re-ranking procedure takes in thousands of these triggers over the entire analysis time, combines them into one file, and assigns the significance via the log-likelihood.

The low-latency analysis assigns the log-likelihood of gravitational wave candidates at the time of the event using the background available at the time. To calculate FAR, these background statistics are marginalized over across all background bins every few hours and written to disk. The most recent marginalized distribution is used at the time of the candidate event to assign a FAR estimation. In the re-ranking procedure, however, the latest cumulative background distribution for each bin, or that which contains background collected over the entire analysis time, is used to rank any trigger. Then, the latest available marginalized distribution across background bins is used to estimate the FAR of that event.

After the significance is re-calculated for all these triggers per background bin, they are merged and reduced across bins. Between any two background bin trigger files, triggers within a coincidence window of eight seconds are compared, and the one with the maximum log-likelihood is kept. The eight second window is chosen to ensure that any triggers which likely correspond to the same candidate are clustered, while also optimally reducing the dataset. We assume that two real gravitational wave events are unlikely to arrive within eight seconds of one another, and use this large window to reduce the number of triggers due to noise in the set. The end result is one file containing the optimal triggers from across background bins in eight second windows for the duration of the analysis. This is then converted to a list of possible gravitational wave candidates similar to those obtained during a full offline re-filtering of the analysis.

The effect of this process is more robust estimation of the significance as it accounts for statistics collected over a longer time span than the original rank assigned in real-time. A trigger collected close to the end of the observing period, for example, is likely to be assigned a similar rank in real-time as the one assigned with the re-ranking procedure because the background used in real-time would contain almost all of the same information as the one used during re-rank. However, a trigger observed early in the observation period would be ranked with a background containing very little of the full observation time, so its rank may vary more significantly compared to the one calculated with this method.

5.3.2 Description of Injection Population

The injections used here follow the distribution of [65], but were made using a new file format, and we will summarize these properties here. The model used to generate the waveforms was IMPRPhenomXPHM [160]. Individual mass ranges were taken between 1-100 $M_{\odot}$, with a limit on the maximum mass ratio given as m_1/m_2 of 20. The primary masses were then distributed using a Salpeter IMF model, and the secondary masses uniformly between the minimum and the primary mass as show in figure 5.1. We consider only spins which are aligned with the angular momentum, with a maximum of 0.2 for each individual mass, and a maximum χ_{eff} of 0.4. Injections are then placed up to a maximum redshift of two, no matter the mass range. The SNR of the waveforms for injections is estimated using the injection parameters, and assuming an IMRPhenomD [161] waveform, the waveform used to generate the templates in our template bank. Then the measured PSD from O3 is used along with a min(max) frequency of 15(1500) to approximate the expected SNR.

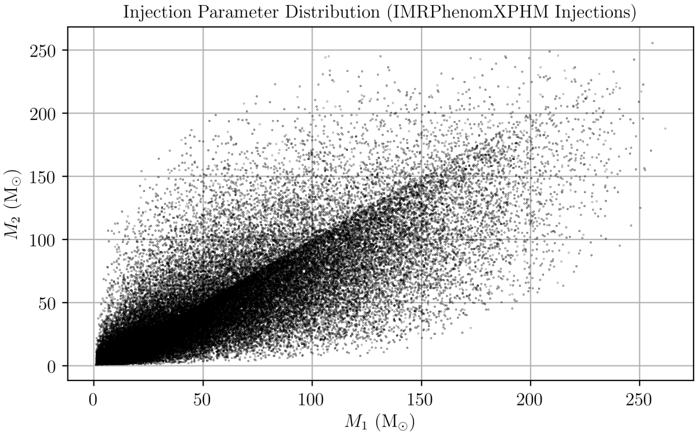


Figure 5.1. Mass distribution of the injected waveforms using a Salpeter IMF mass model for primary masses, followed by a uniform draw of the secondary mass between the minimum allowed mass and the primary mass. Waveforms were made using the IMPRPhenomXPHM waveform model with the minimum(max) individual mass ranges set to $1(100)M_\odot$

.

5.4 Results

In this work, we compare a low-latency analysis over approximately three weeks of O3 data covering January 23rd to Febuary 14th, 2020. This data is replayed as if it were live for analysis by the low-latency pipeline configuration as described in 4.3.1 However, there were intermittent data distribution issues during low-latency operations which caused only a subset of this data to be analyzed. While we can expect small amounts of data distribution problems during any low-latency analysis, typically there are methods to mitigate such issues such as the checkerboard method described in [5]. In this work, we use only one-half of the full template bank, or one checker-boarded bank, and therefore any data outages at the computing center caused full loss of that data.

Under the assumption that the checker-board method would be implemented in a real analysis, any periods of data unavailability to the low-latency analysis which affected the entire analysis, we remove from the data set analyzed by the offline pipelines as well. This ensures that the offline and low-latency analyses are analyzing the same dataset, and therefore most closely comparable. Using this method, we lose a total of 1.25 days in the three weeks of observing time.

It's worth noting that the CAT-1 vetoes applied to the offline analysis are not applied to the online re-rank. A typical re-rank would be completed periodically throughout an observing run, well before CAT-1 veto definitions are available, so we do not apply them here. Additionally, the time segments removed from the analyzed dataset do not also excise the noise and signal background events collected during those times in each analysis. We do not expect, however, that events collected in the 1.25 days should have a large affect the full 23 day distribution, and therefore likely a negligible affect on the sensitivity of the analyses.

The low-latency analysis also suffered from three additional constraints. Due to technical restrictions on uploads to GraceDB servers at the time of the analysis, neither gravitational waves nor injections found by the online analysis were uploaded to GraceDB, and therefore the signal contamination feature described in Chapter 4 was not used. However, this has the added benefit of making online processing more closely resemble that of offline, and reduces the differences between the two.

There was additionally a technical error in a single background bin of this analysis and neither background nor events were collected by this bin. As there are 920 background bins, we do not expect this issue to severely impact the results of this analysis, but this loss may slightly reduce the sensitivity.

Finally, each background bin process in the online analysis writes triggers and noise background information to disk every four hours. If these processes are interrupted or restarted between these four hour snapshots, then all of that information is lost – including the recovered triggers. This kind of intermittent data loss is difficult to account for, but can possibly add up over time to small percentage losses in the sensitivity between the online and offline analyses. This has been fixed ahead of O4b, but the online analysis implemented here suffers from these issues.

All of these caveats are examples of the potential pitfalls that come with real-time analyses. Network issues, computing cluster maintenance, and human error can cause small differences in a real-time analysis compared to what would be expected from a full offline re-filtering where all of these variables are controlled, or correctable. The results we present here then are a true comparison of what might be expected from a typical real-time analysis and the full offline treatment.

5.4.1 Comparison of Re-ranked Low-Latency & Offline Analyses

5.4.1.1 Known Gravitational Wave Events

Here we discuss the gravitational wave candidates which occur during the stretch of data analyzed here, and which were previously published in the GWTC3 catalogue. Of the nine events which fall during the full 40 days of the MDC replay, seven of them occur during the three weeks analyzed by this analysis. The first known gravitational wave events mentioned in [9], GW200112_155838, and GW200115_042309, and the additional retraction occur during the replay before the time covered in this analysis, and so are not discussed here. The results for these seven events can be seen in Table 5.1.

In both analyses, the events GW200202_154313, GW200208_222617, and GW200210_092254 do not meet the FAR threshold of 2 per year for inclusion as gravitational wave candidates, while all other events do. The FAR in the re-ranked analysis is always larger than that of the of fline analysis except for two marginal candidates GW200202_154313 and GW200208_222617 where it is slightly larger in the offline case. However, the gains are small – in all events except for GW200129_065458, the differences in FAR are under an order of magnitude and can be accounted for by the differences in the analyzed data and background distributions, as discussed more fully in the following section.

Note that these results differ from those quoted in [9]. The analysis performed in that work covered a full five week duration of MDC data, while the data here covers a fraction of that time.

Recovered Known GW Events						
		Re-ranked MDC		Offline		
Event Name	SNR	FAR	LnL	FAR	LnL	FAR Ratio
GW200128_022011	9.913	5.356e-13	23.33	3.927e-13	23.14	1.36
GW200129_065458	26.29	2.371e-44	92.01	5.841e-45	93.5	4.06
GW200202_154313	10.87	4.767e-6	8.637	5.15e-6	8.156	0.93
GW200208_130117	9.826	3.988e-13	23.6	3.503e-13	23.26	1.14
GW200208_222617	7.662	1.078e-5	7.657	1.176e-5	7.189	0.92
GW200209_085452	9.652	6.444e-10	17.06	6.332e-10	16.66	1.02
GW200210_092254	8.709	1.864e-5	6.958	1.434e-5	6.923	1.30

Table 5.1. The recovery of known gravitational wave events by the re-ranked online analysis, and the offline analysis over the same period.

Additionally, all of the caveats mentioned at the start of this section apply to the differences here as well. Therefore, it is expected that the sensitivity and recovery of events would differ significantly between the two.

5.4.2 Injection Recovery

In the time covered by this analysis, injections were added every ten seconds in the data resulting in a total of 191,180 injections. However, we do not expect the analyses to recover all of these. For example, only 80,674 (23,539) events have a network SNR of 6(10) and of those, many were not during time when Hanford and Livingston were both active, or those times when we expect our recovery to be best. Additionally, there was an issue with the recovery of low-mass systems in the online analysis for this set, which is an ongoing area of work to rectify, so we will discuss the BBH recovery only. Out of all the injections with a network SNR of 8 or above, there were 13,806 total BBH events where the decisive SNR is given as the SNR in the second most sensitive detector in the network, or the SNR of the only detector operating. This results in 8,264 injections in the chirp mass range of $4.5 - 45 M_\odot$ (low-mass BBH), and 5,542 in the chirp mass range $45 - 450 M_\odot$(high-mass BBH).

The differences in the SNR recovery of these BBH injections for the online and offline analyses can be seen in Figure 5.2 where the label is the standard deviation of the set, and the dashed red lines correspond to ±2 standard deviations. The bulk of this distribution lies well localized around zero with a standard deviation of just 0.004, or 0.4% of the injected SNR. However, there are a few edge cases as seen in the extremes of this figure.

The limits of this data span out to ±50% at the edges. A difference in SNR of 50% could mean the difference between a recovered injection and a missed one, particularly in low SNR cases. An event with an SNR of 3, for example, is not considered a trigger, but an event with an SNR of 4, or a difference of 33%, would be. Only $O(20)$ events live near these extremes, or 0.1% out of over 13,000 total injections in this mass space, but they are unusual for the data set analyzed.

Typically, we would expect differences between an online and offline analysis' SNR

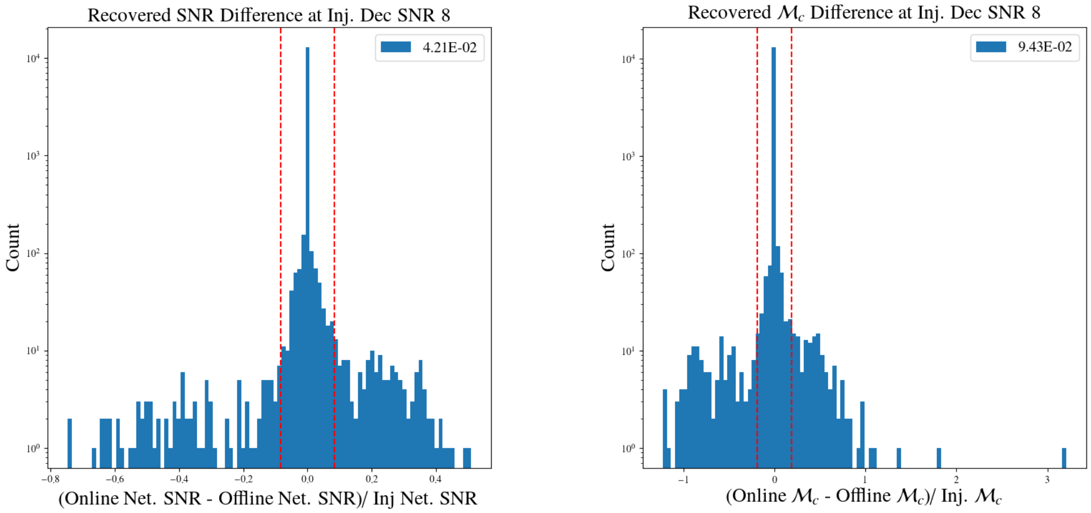


Figure 5.2. Recovery difference of SNR (left) and $\mathcal{M}_c$(right) parameters of approximately 20,000 BBH injections in an online and offline GstLAL analysis. The label denoted the standard deviation and the red dashes denote ± two standard deviations.

recovery due to differences in the data. Ahead of offline analyses, the strain data is re-calibrated, and additional data cleaning is performed in addition to the availability of vetoes. Additionally, in a typical offline and online analysis, the template used may be the same, but they are whitened with different PSDs, and the online template is now re-whitened weekly based on the data. In this test case, however, the online and offline analyses analyzed the same data from O3 embedded in the same noise realization, and the same whitened template bank was used in both to purposefully reduce differences such as these. Therefore any differences in SNR cannot be attributed to differences in data cleaning, or the template bank. Outside of any data distribution anomalies not yet accounted for, then, any differences in SNR must be due to differences in the PSD used to whiten the data as it is collected.

How the GstLAL analysis tracks the PSD is summarized in section 2.1.2. In the low-latency implementation, data is whitened with the PSD available based on only the historical data of the analysis, with information from up to 64 seconds earlier being tracked. In the offline implementation, the data is divided into known segments based on the lock state of the detectors. Then, the PSD is calculated for each segment individually, and the median is used to seed the PSD calculation when performing whitening. Therefore, the PSD used for whitening can vary slightly between the online and offline analyses, especially around the segment boundaries. Typically, we expect this effect to be sub-dominant, but with injections every ten seconds, we are much more likely to catch and probe these most extreme edge cases, and its possible that those most extreme cases seen in Figure 5.2 are created around these segment boundaries.

Differences in the SNR recovery will propagate throughout the rest of the analysis from injection source parameter recovery to FAR estimation, and we can already begin to see its effects in the recovered chirp masses in Figure 5.2 where the label and red dashed lines denote the standard deviation as before. The bulk of this distribution is centered around zero with a standard deviation of 9%, similar to the distribution from a typical offline analysis. It additionally has long tails, however, out to above three at the most extreme end, with the bulk of the tail within one. This shows that while the main distribution is well recovered, the outliers in the online case can become quite large when compared to their offline counterparts, likely a symptom of the outliers in SNR.

Changes in the SNR will also affect the FAR assignment in an non-linear way. Differences in the SNR change likelihood ratio of the recovered injections by altering several terms in both the noise and signal models which are dependent on the SNR. The signal and noise background histograms used by the $\ln P(\mathrm{SNR}, \xi^2/\mathrm{SNR}^2)$ likelihood ratio term, for example, are both parameterized in part by the SNR and these will be affected by this recovery. A simple change in the SNR will change the distributions of these histograms, and if the SNR is particularly smaller then normal, some events which would go into the background could also be missing entirely. Both of these effects will shift the parameter space covered by the noise and signal distributions, and therefore change the likelihood ratio.

While changes in the SNR will change the noise background distributions, the timeslide used to generate background samples in the offline analysis will cause entirely unique samples to be added in the offline case. In the offline analysis implementation, we perform a single time-slide operation which generates these samples. This operation takes foreground triggers, or those with a net SNR above 7, in Livingston and Virgo and shifts them just once by 66% and 33%, respectively. The time-shifted sets are then used to form new coincidences between one another and the original Hanford foreground samples. This effectively forms false coincidences between foreground triggers and noise triggers in the detectors, giving a measure of whether your foreground model would match the noise model if paired with random noise.

If the network SNR of these coincidences is less than 7, but the individual detector SNR is greater than 4, then a new sample is added to the background distribution corresponding to this trigger with the new coincidence's ranking. The false coincidences will retain the same properties as the original noise triggers in each detector such as SNR and ξ^2, but will now have a different log-likelihood as the log-likelihood is based on the combined parameters of the triggers in each detector, and the properties of the coincidence itself. The samples generated for the noise distribution in this method are then kept when assigning significances to the original data, thereby changing the samples present in the background, as well as increasing the number of total samples. An example of how this changes the background distributions for a single bin can be seen in Figure 5.3.

The additional samples from the timeslide alone will cause a change in distribution of the background thereby affecting the SNR, ξ^2 term assigned to the noise and signal models as a result. However, the signal and noise histograms also both have a KDE applied before use in the likelihood ratio. The KDE implemented follows Silverman's rule which tunes the bandwidth based on the number of samples in the set. Therefore, a change in the total number of samples will cause the KDE bandwidth to be tuned to be smaller offline than online, which will further affect the likelihood ratio assigned to events. As the FAR depends both on the likelihood ratio assigned to the event and the bulk of likelihood ratios assigned to noise events through the CDF, even small differences in likelihood ratio assignment and background propagate to the FAR calculation.

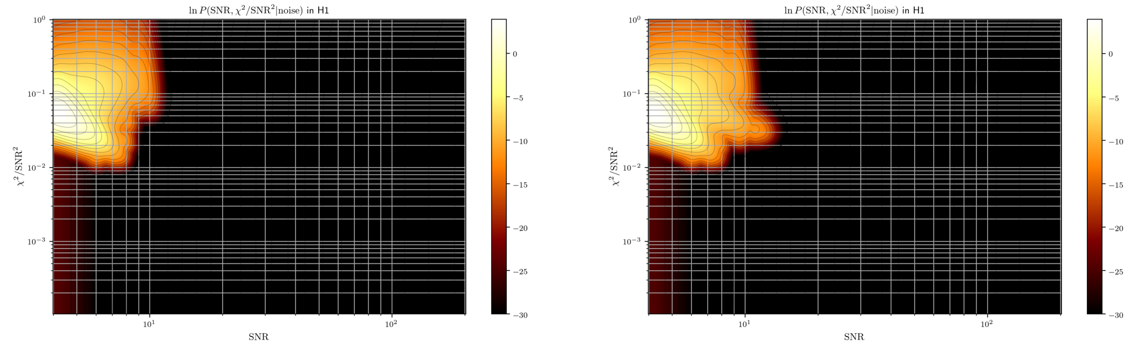


Figure 5.3. left: The noise model distribution of the lnP(SNR, χ^2/SNR^2|noise) term in the first background bin in H1 for an online analysis. right: The same as in the left figure, except for an offline analysis.

Although we saw few changes between the FAR of the gravitational wave events in this dataset due to these changes, the injections probe the full depth of the differences between online and offline as shown in Figure 5.4. The left figure shows a limited axis in order to highlight the features around zero, while the right figure shows more of the total distribution with about 1% of the tail not shown for clarity, as it extends with outliers out to past 1e25. This long tail shows that offline events are more frequently recovered with a lower FAR than the online events, as was expected with a few particularly extreme samples. However, the median of the set is 1.12, or a difference of two orders of magnitude between the online and offline results for the bulk of the distribution.

5.5 Search Sensitivity

The differences in FAR from the previous section will have a large affect on the sensitivity of the online pipeline as a whole. If the FARs are systematically higher, then fewer injections will be recovered above the typical FAR thresholds, and the pipeline will suffer in both efficiency and $\langle VT \rangle$.

The efficiency is a measure of how well the pipeline recovers injections, and is given as:

$$\epsilon(FAR, \mathcal{M}) = \frac{N_{\text{found}}(FAR, \mathcal{M})}{N_{\text{total}}(\mathcal{M})} \tag{5.1}$$

where $\mathcal{M}$ represents the dependency on the chirp mass range. The results of the offline and online recovery using this measure can be seen in Table 5.2. There is a natural decrease in efficiency with increasing FAR threshold, and an order one percent loss in the online analysis across FAR thresholds. This loss is in part due to the skewed FAR recovery seen in the previous section, however, it could also be due to differences in observing time between the analyses.

As discussed in previously, each of the online analysis bins can lose triggers if the process happens to be interrupted. These can add up over time to a percentage loss in observing time as seen in Figure 5.5. Here, the number of bins out of 912 for each interferometer is counted on the y-axis, and the percent of time missing from that bin compared to the total time observed

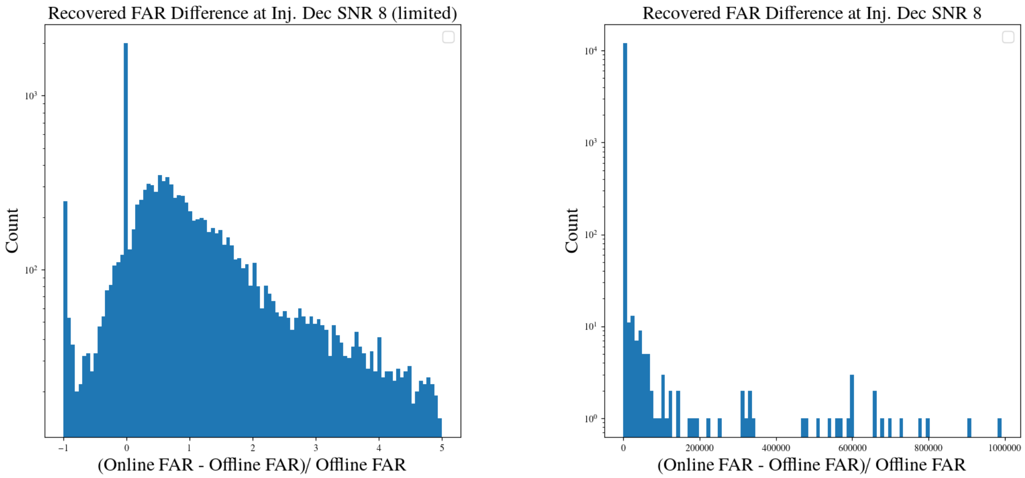


Figure 5.4. The false alarm rate (FAR) estimate differences between an online and offline analysis analyzing approximately 20,000 BBH injections over the course of 23 days. Left: A subset of the distribution with about 25% of the total distribution cut out of the tail. Right: The full distribution with 1% of the tail not displayed.

by the offline analysis is shown on the x-axis. For each online bin process, there is anywhere between 0.5 and 2% missing on average.

The union of covered segments across bins is used when calculating the number of injections which are recoverable, so these small differences will not be accounted for the total number of injections, but can cause a loss in the number actually recovered. That is, if 911 of the 912 bin processes were not collecting data, then all injections in that time would be counted in N_{total} despite it being highly unlikely that a single bin would recover all injections. This is an extreme example, and typically not more than a handful of bins are ever interrupted simultaneously except during cluster maintenance times, but clusters of bin interruptions could affect recovery, particularly with the frequency of injections used here.

	Re-ranked MDC		Offline	
FAR	low-mass BBH	high-mass BBH	low-mass BBH	high-mass BBH
2.78e-4	0.927	0.900	0.935	0.905
2.31e-05	0.904	0.876	0.913	0.883
3.85e-07	0.866	0.834	0.875	0.839
3.16e-08	0.840	0.804	0.849	0.807

Table 5.2. Efficiency of the re-ranked online and offline analyses using a decisive SNR threshold of 8 on a full population of over 13,000 BBH injections using multiple FAR thresholds to consider these injections recovered. The FAR thresholds correspond to one per hour (the GraceDB upload threshold), two per day (the public alert threshold), one per month, and two per year. The low-mass range described here accounts for masses in $4.5 - 45 M_{\odot}$, while the high-mass range is $45 - 450 M_{\odot}$.

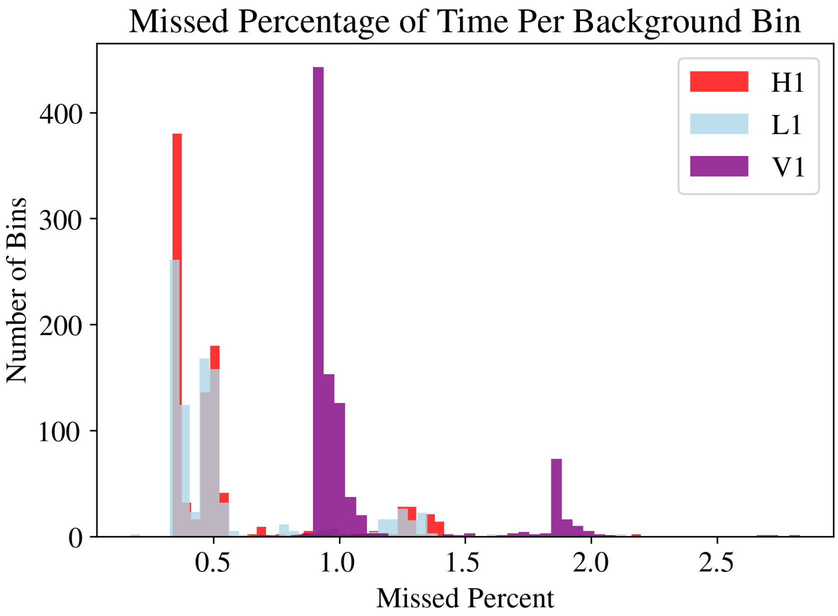


Figure 5.5. The amount of time missing from individual background bin processes compared to the total time observed in each ifo over the course of 23 days.

The differences in efficiency between the two analyses go on to have a direct impact on the measured $\langle VT \rangle$ as seen in Figure 5.6. This figure shows the $\langle VT \rangle$ ratio between the online and offline analyses for the BBH mass ranges where the $\langle VT \rangle$ is given by Equation 4.5. The $\langle VT \rangle$ ratio between the two analyses in these mass ranges is similar to the loss in efficiency at those FAR values, but is typically higher than from the efficiency alone. Notably, the online analysis occasionally outperforms the offline analysis, particularly at the smaller mass range. Although the efficiency does not account for this improvement over offline, the $\langle VT \rangle$ also depends on the luminosity distances of the sources which each analysis recovered. The two analyses will have recovered a slightly different subset of all of the injections which causes different weightings by the luminosity distances. The different luminosity distances will result in a varying $\langle VT \rangle$ compared to the efficiency alone and likely accounts for the few percent differences shown when taking into account thousands of injections.

When considering the computational cost and person-power required for offline analyses, a $\mathcal{O}(1\%)\langle VT \rangle$ loss at high FAR is an acceptable one. This reinforces just how minute the differences between these analyses are, and that small improvements the offline analysis gains in SNR, and FAR account for a very small number of injections in the larger picture. With all of the caveats in the online analysis discussed here, however, this result still has room for improvement. The changes made ahead of O4b to rectify the individual bin losses, as well as the implementation of a checker-boarded analysis will both increase the overall uptime of the online analysis and the availability of accurate scientific output. In addition, improvements can be made to the background collection method of the online analysis. Collecting more background samples would of course increase the sensitivity, but other methods for tuning the

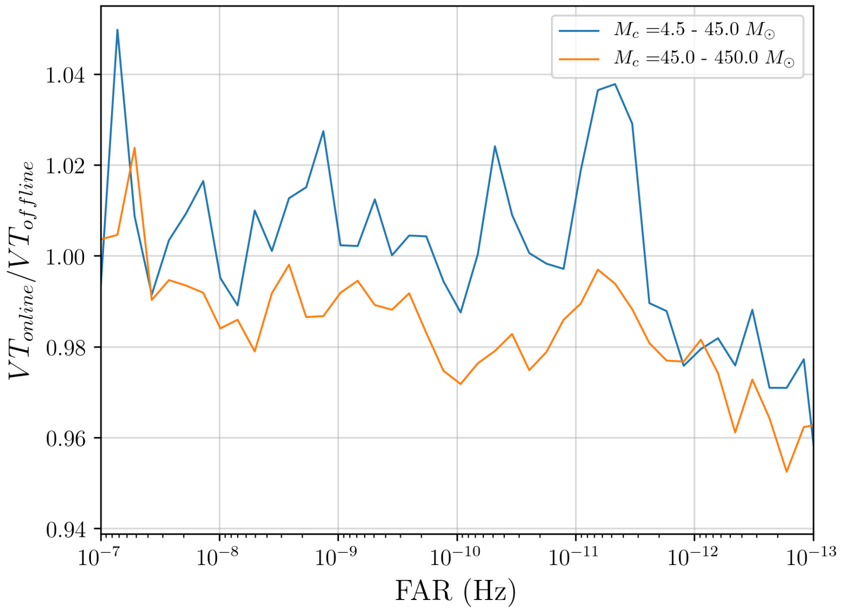


Figure 5.6. $\langle VT \rangle$ ratio of the online and offline analysis sensitivities in two BBH mass ranges.

bandwidth estimation in the KDE could also be investigated. While this is an open area of research, these initial results are promising and demonstrate the the near-future attainability of the exclusive low-latency analysis of gravitational waves.

Chapter 6
Data Quality in Gravitational Wave Interferometers with iDQ

6.1 Noise Transient Searches

6.1.1 Background

For any gravitational wave signal, the strain and amplitude are defined in the time domain as in Chapter 1. Of interest for short duration noise transient detection pipelines is additionally the signal in the frequency domain given by:

$$\tilde{h}(f, \mathcal{A}) = \mathcal{A}\tilde{h}(f) \tag{6.1}$$

where $\tilde{h}(f)$ represents the Fourier transform. Then, the characteristic time, t_0, and characteristic frequency, f_0 can be expressed as the expectation values:

$$t_0 = \int_{-\infty}^{\infty} t|h(t)|^2 dt$$
$$f_0 = \int_{-\infty}^{\infty} f|\tilde{h}(f)|^2 df$$

The bandwidth of the signal is calculated using the variance in frequency as:

$$\sigma_f = \int_{-\infty}^{\infty} (f - f_0)|\tilde{h}(f)|^2 df \tag{6.2}$$

Q, or the quality factor of a signal, is then the ratio of the characteristic frequency to the bandwidth:

$$Q = \frac{f_0}{\sigma_f} \tag{6.3}$$

Signals with high Q have a well defined bandwidth and therefore low uncertainty in their frequency, or many oscillations per envelope and are therefore narrow-banded. Meanwhile, low Q signals have a high uncertainty in their frequency and few oscillations per envelope.

6.1.2 Omciron

Omicron [29] uses the Q transform of the time-frequency space around signals to identify regions of excess SNR, and therefore potential glitches. They use sine-Gaussian signals to define a basis and apply a Bisquare window to limit the sine-gaussians in time from the true infinity of gaussian distributions. The resulting window distribution used is defined as [30]:

$$\tilde{w}(f, f_0, Q) = \begin{cases} \mathcal{N}\left(1 - \left(\frac{fQ}{\sqrt{11} f_0}\right)^2\right)^2 & f > \frac{f_0\sqrt{11}}{Q}, \\ 0 & \text{otherwise} \end{cases} \tag{6.4}$$

where $\mathcal{N}$ is a normalization factor, Q is the quality factor, f_0 is the characteristic frequency defined above, and the factor of $\sqrt{11}$ appears here from use of the Bisquare window instead of a true gaussian. Then, the Q transformation is performed by projecting the known signal onto the sinusoidal basis per tile where one tile corresponds to one basis sinusoid with a given central frequency, central time, and Q. However, signals are limited in their uncertainties by the classic relation:

$$\sigma_f \sigma_t >= \frac{1}{4\pi} \tag{6.5}$$

This results in low Q tiles having a small duration and wide bandwidth while high Q tiles have a long duration and small bandwidth. The projection for a single tile, X, is done as follows [30]:

$$X(\tau, f_c, Q) = \int_{-\infty}^{\infty} x(t) w(t - \tau, f_c, Q) e^{-2\pi i f_c t} dt \tag{6.6}$$

where τ is the central time of the tile, f_c is the central frequency of the tile, $x(t)$ is the timeseries of the signal, and w is the window function in time. The projection is then cast into the frequency domain as [30]:

$$X(\tau, f_c, Q) = \int_{-\infty}^{\infty} \tilde{x}(f + f_c) \tilde{w}^*(f, f_c, Q) e^{2\pi i f \tau} df \tag{6.7}$$

where the window $\tilde{w}$ can then be taken from 6.4 and $\tilde{x}$ is the Fourier transform of the timeseries. Then, for an entire set of tiles covering the full time-frequency plane of a given Q, f_c and τ must be iterated over for each tile in the plane. The tiling distribution used by Omicron in a single Q plane is defined by an acceptable energy loss due to mismatch between the tiles. This tiling strategy leads to sets of tiles defined logarithmically in central frequency and Q, and linearly in central time [30].

With this, the excess energy in any given tile in the Q transform, $X(\tau, f_c, Q)$, as the ratio of the energy in that tile to the mean expected energy of all tiles in a given Q plane,$\langle X(Q)^2 \rangle$ [30]:

$$\rho^2 = \frac{|X(\tau, f_c, Q)|^2}{\langle X(Q)^2 \rangle} - 1 \tag{6.8}$$

$$\rho^2 = |X(\tau, f_c, Q)|^2 - 2 \tag{6.9}$$

where the subtraction of one results in tiles which match the expectation giving an excess power of zero, and the expectation value for white noise is 2 giving the latter result. This value is taken as an estimate of the SNR, ρ.

When the excess power is defined for every tile, Omicron then down-selects tiles to produce triggers by keeping only those tiles which have an SNR of 5.5 or greater. Omicron then further clusters these triggers into events as there could be triggers from many tiles corresponding to the same glitch or gravitational wave. This is done by clustering all triggers, despite frequency or Q value, in a rolling 0.1 second window. That is, any down-selected trigger and its neighbor are considered to correspond to the same event if their central times are within 0.1 seconds of one another. This can result in a varying duration of a single cluster, taken as the difference between the minimum and maximum central time of triggers in that cluster. For example, if there are only two nearby triggers, then the cluster duration would be 0.1 seconds, but if there are many, this could increase indefinitely. In reality, the longest known clusters are up to order 10 seconds in length.

For each of these clusters, the peak time, peak frequency, and SNR of the cluster is assigned as those from the highest SNR tile in the cluster. Then each cluster is assigned a start time, or the central time the earliest tile in the cluster, an end time, or the central time of the latest tile in the cluster, SNR, peak frequency, peak time, the frequency limits of tiles in the cluster, and finally additional information about the highest SNR tile in the cluster. These clusters are added to the Omicron database periodically when being generated offline, or continuously when generated automatically online. In Omicron's online configuration, this clustering process combined with the runtime of the package incurs an order thirty second latency compared to real time data. Any outages of Omicron online are manually back-filled to keep full coverage of available observing time.

6.1.3 SNAX

SNAX [31] is a toolkit inspired by Omicron, but with the added benefit of processing data in near real-time for use by downstream processes such as iDQ.

Similarly to Omicron, SNAX builds a basis on windowed sinusoids, but implements a Gaussian window. These basis waveforms can be expressed as:

$$h(t) = Aw(t - \tau, f, Q)\cos(2\pi f t + \phi) \tag{6.10}$$

where A is the normalized amplitude, τ is the central time, Q is the quality factor as defined previously, f_c is the central frequency, ϕ is the phase, and w is the Gaussian window. In this case, a typical Gaussian window is used with a tapered acausal component to minimize latency defined by [31]:

$$w(t - \tau, f_c, Q) = \begin{cases} exp\left(-\frac{1}{2\sigma_t^2}(t-\tau)^2\right) & t < 0 \\ exp\left(-\frac{1}{2\sigma_t^2}(t-\tau)^2\right) exp\left(\frac{log(\epsilon)(t-\tau)}{l_{max}}\right) & \end{cases} \tag{6.11}$$

where l_{max} is the maximum desired incurred latency from the matched filter, ϵ is a pre-determined tolerance for truncation on the Gaussian waveform, given as 5e - 3 in a typical analysis, and σ_t^2 is the variance on time given by:

$$\sigma_t^2 = \frac{Q^2}{8\pi^2 f_c^2} \tag{6.12}$$

Thus, a template bank of the sine-gaussian basis vectors is constructed and parameterized on central frequency, f_c and the quality factor, Q. SNAX analyzes thousands of auxiliary channels in addition to the strain channel, so templates are placed in a bank per auxiliary channel subsystem in a frequency range appropriate to that sub-system. In the general use case, they are alternatively placed by default in the gravitational wave search frequency range and in a Q range with a minimum set for anti-aliasing effects and a maximum based on computing constraints.

Using these templates, each target channel's timeseries is resampled, whitened, and then matched-filtered with the corresponding template bank. The output is an SNR timeseries per template in the bank which is then aggregated across templates, similarly to the GstLAL analysis, by keeping the template and SNR corresponding to the highest SNR in some small window, t_w. This results in one SNR timeseries per desired channel along with information about the template parameters per t_w. Then, the results are bundled and synchronized across channels, so the output contains results from all channels of interest across some window of time, typically one second in low-latency operations. These synchronized series are then ingested by downstream processes such as iDQ.

6.1.4 GravitySpy

While other detector characterization pipelines like Omicron contribute by identifying glitches in the timeseries strain data, GravitySpy contributes classifications to those times. As discussed in section 1.4, there are variety of glitch classes defined by their morphology in the time frequency space. GravitySpy uses a CNN machine learning algorithm [27] to classify a time of interest based on spectrograms of the time with four durations: 0.5s, 1s, 2s, and 4s. The variety of durations is required to expose the morphology of glitch classes with varying duration. The output of GravitySpy for a time of interest is a list of all glitch classes considered each with an assigned probability, or confidence. The confidence across classes is not required to sum to one, and instead each class is assigned a value between 0 and 1 individually. For example, GravitySpy could be confident that there was a scattering-like glitch present, but be unable to distinguish whether it was a Fast Scattering glitch or Scattered Light. In this case, you could expect high confidence values in those two classes and lower confidence values across all others.

The training set for GravitySpy has evolved with time through observing runs as more glitches are discovered, and more morphological classes are defined. As of LIGO's third observing run, the training data consisted of 9631 labeled glitch samples across 23 morphologies. Of note, one class included is the "No Glitch" class which was trained to correspond to times when there was no glitch present. However, as discussed in more detail in Chapter 7, this class is frequently assigned by GravitySpy to glitch glitches which don't match any of the other known morphologies – even at high confidence.

During active LIGO observing runs, GravitySpy classification is triggered on new Omicron event times uploaded to the database with SNR of 7.5 or greater. This allows for medium-latency identification and classification of glitches which informs detector maintenance.

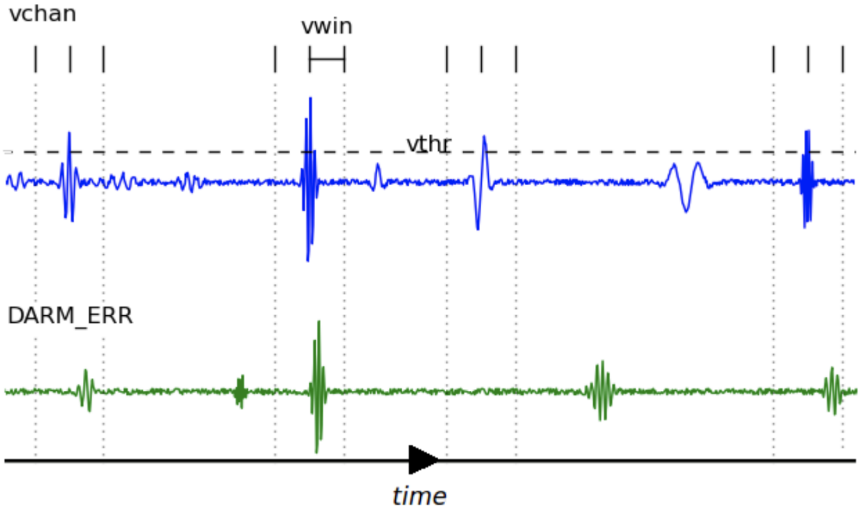


Figure 6.1. Cartoon example of how OVL identifies and counts correlations between an auxiliary channel (top), and the strain channel (bottom). For each auxiliary channel (vchan), a window (vwin) and threshold (vthr) is assigned to create a veto. A variety of statistics are then calculated based on how these two combine to remove glitches from the strain data. Reprinted from [10]

6.1.5 OVL

OVL [10] identifies correlations between transients in the strain data, and auxiliary channels by constructing an Ordered Veto List. To do so, OVL leverages the use of a labeled dataset. By default, it uses the Kleine-Welle algorithm to label a given dataset, but in theory any glitch finding algorithm can be used which provides a time and significance estimate. For example, when implemented in iDQ, it is supplied feature vectors from SNAX or Omicron containing time, frequency, and SNR.

OVL then searches for correlations by setting windows and thresholds on transients in the auxiliary channel data and searching for coincidences in the strain data. A cartoon example of this is shown in Figure 6.1.

A variety of windows and thresholds are set on the auxiliary channels to account for the varying morphology of glitch types and classes. A veto in an Ordered Veto List is then defined by three characteristics: the auxiliary channel, the window, and the threshold. Segments representing the time removed by application of the veto are constructed by applying the window around any time that crosses the threshold. In this way, the dead-time, t_d, introduced by application of a veto is simply a sum of these segments, and the fractional dead-time, f_d, is the dead-time over the total livetime observed, or $f_d = t_d/T$. The efficiency, ϵ of the veto is calculated as $\epsilon = n_c/N_{\text{strain}}$ where n_c is the number of coincidences between the strain data and the auxiliary channel transients above the threshold, and N_{strain} is the total number of transients in the strain data. Assuming that glitches are generated by Poisson processes, the efficiency over dead-time ratio for each veto is calculated as follows [10]:

$$\epsilon/f_d = \frac{n_c/N_{\text{strain}}}{t_d/T} \tag{6.13}$$

$$= \frac{n_c}{t_d(N_{\text{strain}}/T)} \tag{6.14}$$

$$= \frac{n_c}{t_d\lambda_{\text{strain}}} \tag{6.15}$$

where λ_{strain} is the expected rate of glitches in the strain data for a Poisson process, and T is the total observing time. This is the number of coincidences observed over the expected number from a Poisson process and is used the main measure of significance in OVL. Its worth noting that this measure can be large even with a small number of coincidences identified if the deadtime introduced is also small. For example, if an auxiliary channel is an excellent witness for a single rare glitch type, then n_c may be small, but if it witnesses these glitches precisely, t_d could also be small and still form coincidences. In this case, the efficiency-deadtime ratio may not vanish even though the channel witnesses only a small number of coincident events.

The efficiency-deadtime ratio is calculated for each of the vetoes using all of the time in the dataset as an initial significance estimation. Vetos are then ordered from largest to smallest significance to construct the first ordered-veto-list iteration. Then, the vetoes are applied to the original dataset one-by-one from the top of the list to the bottom and the significance calculated where the time covered by each subsequent veto is removed from the dataset before applying the next. In this way, each veto is actually being applied to a different dataset, but the significance of any redundant vetoes goes to zero. After ordering, the list is pruned by a set threshold on the efficiency-deadtime ratio, and the Poisson significance. The Poisson significance is taken as the probability of observing as many glitches at the Poisson expected rate, λ_{strain} or more and is given as [10]:

$$p = \sum_{k=n_c}^{\infty} \frac{\langle n_c \rangle^k}{k!} \exp\left(-\langle n_c \rangle\right) \tag{6.16}$$

$$= \sum_{k=n_c}^{\infty} \frac{(n_c f_d/\epsilon)^k}{k!} \exp\left(-n_c f_d/\epsilon\right) \tag{6.17}$$

where $\langle n_c \rangle = t_d\lambda_{\text{strain}}$ is given as the expected number of coincident counts from a Poisson process. Note that this is not the same as the efficiency-deadtime ratio and prefers vetoes with large coincident numbers at a given efficiency-deadtime ratio. By requiring vetoes to pass both the efficiency-deadtime ratio and Poisson significance threshold during pruning, OVL requires both correlation strength and large coincident numbers. As mentioned previously, requiring only high efficiency-deadtime ratio can allow for veto configurations with small numbers of coincidences and small deadtimes to be kept. After initial ordering and first pruning, the vetoes are then re-ordered by the calculated significance and the process is repeated until the list converges. The act of pruning allows for this process to converge more rapidly. In this way, correlations between any number of auxiliary channels and any strain dataset can be identified and ranked by their significance.

6.2 iDQ Methodology

iDQ identifies non-Gaussian noise via auxiliary channels, and provides probabilistic statements on the likelihood of evaluated strain data containing a glitch. Full details on iDQ are given in [11]. An overview is provided in the following section following that work along with additional detail on how this methodology changes between the online and offline implementations.

iDQ uses a two-class classification scheme to define glitches, and any time that is not a glitch, or is clean, such that:

$$p(G) + p(C) = 1 \ \forall t \tag{6.18}$$

Additionally, iDQ does not further classify glitch times like GravitySpy, instead just assigning them to the generic glitch class. However, as will be discussed in section 7.4.1.1, the classifications of GravitySpy can be combined with the results of iDQ to both further classify the times iDQ identifies, and extend this to trace glitches back to their auxiliary channel sources.

Labels for iDQ's training set are derived from the strain data via SNR thresholds on Omicron and SNAX triggers as described in detail in 6.2.2. The determination of classification by iDQ, however, is dependent only on the auxiliary channels in each interferometer. iDQ analyzes only the safe auxiliary channels in each, or those that are insensitive to genuine gravitational wave signals. This prevents iDQ from identifying gravitational waves, making it ideal for incorporation into gravitational wave detection pipelines as a veto or similar data quality flag. How iDQ processes safe auxiliary channels to assign the glitch classification is discussed throughout the duration of this section.

6.2.1 Data Discovery and Division

The main input to iDQ is a set of features represented by tabular data on auxiliary channels. iDQ could, in theory, ingest the raw detector data, but using a feature extractor allows for upstream data handling, and transformation thereby simplifying the iDQ pipeline. There are a variety of feature extractors which each report their own feature sets using their own methodology. Omicron, see 6.1.2, for example uses Q transformed tiles to estimate SNR, frequency, central time, and duration of noise transients in the target channel. Whereas SNAX, see 6.1.3, uses matched filtering to detect noise transients and estimate the SNR, frequency, and time. Offline production of iDQ currently implements Omicron, while online production implements SNAX as the upstream feature extractor.

The cadence at which feature extractors report features on individual channels may change with time, and there could be too many or too few features for iDQ to use in the analysis. Therefore, ahead of ingesting the data, iDQ converts features from the upstream feature extractor into feature vectors. This is done by taking the features for any single channel within some window (typically about 100ms) and taking the maximum SNR feature in the set, or reporting default values if no feature is available. This allows for a consistent dimensional input to the pipeline at all times.

With this, there are two methods for dividing data for training and evaluation: acausal and causal. The acausal scheme is used exclusively for the offline iDQ analysis and can be seen in Figure 6.2.

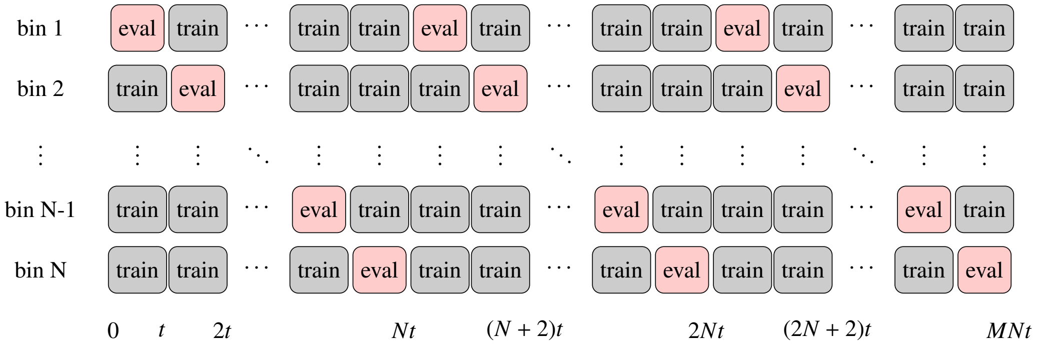


Figure 6.2. Schematic diagram showing data segmentation for acausal batch operation. Each row corresponds to one of N bins, and each column corresponds to one of $N \times M$ segments. Across all bins, only one segment is used for evaluation, but across bins they add to cover the full time of interest. Reprinted from [11].

It first chooses N bins into which to divide the data with M segments in each bin. Then, the entire time of interest is divided into N x M segments. The first segment is assigned to the first bin, the second to the second, and so on until the (N+1)segment is again assigned to the first bin, the (N+2)segment to the second, etc. This process is repeated until each bin has M segments in it. The Nth segment of all M segments in a bin is then used for training while all others are used for evaluation. This system creates N bins each with information from a wide distribution of the total observing time for use in training. Additionally, across each of the N bins for any time segment of interest, there is only one bin which uses that time segment in evaluation, resulting in a unique and continuous output timeseries. It does require, however, that the entire duration of interest be available at the time of analysis which is why it lends itself exclusively to offline, and high latency usage.

The causal method, however, does not have such constraints. In this method, all data before some recent history is used as training data for the current time cumulatively. This requires an initial look-back in order to collect the data for first training set, but afterwards incurs no additional latency and is constantly updated with the latest information from the dataset. This works well for real-time analysis as any changes in the detector characteristics are consistently folded into the analysis at with a cadence equal to the segment length, typically 3 hours.

6.2.2 Training

In order to evaluate a desired dataset, iDQ must first train a classifier to identify glitch times. Although iDQ is constructed such that any classifier can be used, OVL has been the classifier used exclusively up to this point for production-level iDQ analyses, and this implementation will be discussed in the most detail here. Training datasets for iDQ consist of labeled feature vectors from times chosen via the binning scheme discussed in the previous section. The labels for these vectors are assigned based on a threshold on the significance of the vector. In a typical production analysis, the glitch label is applied to vectors with an SNR of 10 or greater, while the clean label is applied to those with an SNR of 5.5 or less. Vectors with significance between the two labels are considered dirty, and are not used in training. Additionally, clean vectors are required to be a full second away from any glitch vector thereby ensuring independence of any two samples.

A classifier takes these labeled vectors and generates a model, or a mapping from the input parameter space of any arbitrary feature vector to a rank space between zero and one. The details of each mapping is unique to each classifier and which of the input feature vector parameters are used and considered relevant is entirely unique to each. Additionally, the chosen rank space between zero and one is an arbitrary one made for iDQ. What matters is the ordering of the results within the rank space, and the range of zero to one was chosen for convenience while any range would contain the same information. This mapping from input parameter space to rank space is then stored as a trained model, and given a unique hash for usage in evaluation later in the pipeline.

In its application within iDQ, the OVL classifier has been modified to use the efficiency-deadtime ratio, the Poisson significance, or the use percentage to order the vetoes in its lists. The efficiency-deadtime ratio and Poisson significance are as described in section 6.1.5, and the use percentage is the number of coincident glitch events identified at a certain threshold over the total number of transients in the auxiliary channel which pass that threshold, or $U = n_c/N_{aux}$. Whichever metric is chosen, the value of the metric is mapped from the possible [0,inf) space to the [0,1] space of the rank, R_m, via a scaling factor as [11]:

$$R_m = \frac{s_m * x_m}{1 + s_m * x_m} \tag{6.19}$$

where s_m is the scaling factor for the metric and x_m was the value of the metric itself. No matter the metric chosen, the rank is now used to order the vetoes in lists. Pruning of the lists is designed to allow enforcement of a minima on any of the statistics which are available for use in the rank. In the production iDQ online analysis, for example, minima are set on Poisson significance, use percentage, and the efficiency-deadtime ratio as 5, 0.05, and 5 respectively while the rank is based solely on efficiency-deadtime ratio. However, as will be discussed in section 7.3 changes to the available ranks were made ahead of the fourth observing run to allow for a wider variety of considerations.

6.2.3 Evaluation

After training, the models are then available for application to data of interest, as assigned in section 6.2.1. During the evaluation phase, the models made from the training data are applied to validation data and the result is stored, along with a reference to the original model hash. In offline running, all of the data for evaluation per bin is loaded in at once and has the single model for that bin applied. This process is repeated across bins resulting in an evaluated dataset for every time in the input range. During online running, a stride's worth of data (typically about 900s) is loaded in, the most recent model is applied, and the result is written to disk as a ranked dataset. Then this process is constantly repeated in a cadence equal to the stride length.

To apply the model, the feature vectors for each channel are compared to the ordered veto list stored in the model. In offline operation, the model used here is the one created by the bin in question, while in online it is the latest model. The vetoes in the OVL are applied in order as described in 6.1.5, and the result is a rank value for every time in the stride.

Note that the output of this evaluation process is not the deliverable timeseries data, but it is instead classified datasets of a stride's length. These resulting datasets are passed on to calibration map processes for use in calibrating the log-likelihood and FAP, but are not reported as the output timeseries. As discussed in section 6.2.5, artifacts in timeseries generation can cause the ranks calculated here to mismatch those reported by the actual timeseries process.

6.2.4 Calibration

The calibration processes form maps from rank to probabilistic statements, called calibration maps. While the rank could be the only output of iDQ, probabilistic statements are not only more human-friendly, but can be implemented directly in downstream gravitational wave detection pipeline ranking statistics. The timeseries which iDQ currently calibrates are a log-likelihood, and a false alarm probability, although there will be discussion on a glitch probability statement in section 7.4.1.1. The likelihood-ratio is given via the Neyman-Pearson lemma as:

$$\Lambda_C^G = \frac{p(\text{data}|G)}{p(\text{data}|C)} \tag{6.20}$$

where $p(\text{data}|G)$ is the probability of getting the data given that the time given has a glitch label, and $p(\text{data}|C)$ is the same for clean labels. To calculate each of these probabilities, clean and glitch times are taken from the training and evaluated datasets. All of the glitch samples are taken and used to generate a histogram of glitch ranks, while only a random sampling of the clean times is taken as there are frequently two orders of magnitude more clean samples available than glitch samples. Probability density functions (PDF) of these histograms are modeled using a Gaussian kernel density estimate (KDE) with reflected values at the rank =0/1 boundaries to avoid edge effects. The log-likelihood is calculated from these PDFs as the ratio of the log of the glitch PDF over the log of the clean PDF at a given rank value.

This same clean PDF is then also used to calculate the false alarm probability (FAP) statistic taken as [11]:

$$\text{FAP} = \int_r^1 p(M(\text{data}) >= x|C)dx \tag{6.21}$$

where M(data) is the output rank of the model on the data and therefore $p(M(\text{data}) >= x|C)$ is the probability of the rank assigned by the model being greater than the rank threshold given that the data is not a glitch. To obtain this, the PDF of the clean samples is integrated cumulatively to form a cumulative distribution function (CDF). The FAP can then be taken as FAP = 1 - CDF(r), or one minus the CDF at a given rank to obtain the integration from that rank to 1.

Thus, any calibration map contains the PDF of the clean and glitch samples from the training data up to that point as well as the CDF of the clean data. It can then be applied to any time of interest in order to immediately calculate the calibrated log-likelihood and FAP. In offline production, this calibration map only needs to be calculated once per segment, like the evaluated data. In online mode, the calibration map is updated as soon as new evaluated data is available, and is calculated from scratch when new models are available. It is therefore dependent on the training and evaluated data cadences, but is typically updated every few minutes, and calculated anew every few hours.

6.2.5 Timeseries

For any given time of interest, iDQ applies the latest model and calibration map to obtain the rank, log-likelihood, and false alarm probability through the timeseries processes of the analysis. Timeseries processes poll for feature vectors in a stride's length, one second in online processing and 1800 seconds in offline processing. Models are applied to this data in the same manner as the evaluation process as described in 6.2.3 to obtain a rank timeseries. These rank values are then converted to log-likehood and FAP by applying the calibration maps calculated as described in section 6.2.4. The resulting timeseries of all three are written to disk for the stride length in question as the main output of iDQ. Then, a new stride's worth of data is loaded and the process is repeated.

One weakness of this process is revealed in the online operating mode which uses the short stride of one second in order to keep up with real time. However, information about the veto segments and windows are not kept between strides. That is, if a veto with a half second window is active at time 0.99s, the analysis does not record that it should be marked as active at time 1.01s. Similarly if it is active at 1.01s, it does not retro-actively apply the veto to 0.99s. The latter follows because these times have already been written to disk, and in the online mode, distributed to downstream users. However, discarding the information before being applied to future times is a shortcoming of the current configuration which can cause a mis-match between the information represented in the timeseries and the information represented in the datasets output by the evaluation processes. The evaluation processes are set with a significantly longer stride than that of the timeseries procceses, meaning this lack of memory affects the timeseries more significantly than the evaluated datasets. The impact is that the classified datasets made by the evaluation processes differ from the ranked timeseries on disk.

This effect is mitigated in offline processing as a significantly longer stride (about 1800s) is used to calculate the timeseries, so these edge effects will be much more rare. Mitigation of this issue online is left to future work, and the speed and computing performance of the analysis will have to balanced with the benefits of keeping longer stretches of data in memory.

6.2.6 Online vs offline running

Although discussed intermittently in the previous sections, we collate the many differences in the iDQ process when run in the online, low-latency configuration versus in the offline, high-latency configuration here. In these two modes, the biggest difference is the latency incurred and the cost of that latency.

In the most common configuration of offline mode, Omicron is used to create input feature vectors, and the data is binned using the acausal scheme described in section 6.2.1. This necessitates the availability of data covering the entire time range of interest from the start of the analysis, and therefore incurs a minimum latency of the duration of the time of interest. It also, however, accounts for data from across the observing time of interest instead of only the data up to the time analyzed. For example, the training data is taken as shown in Figure 6.2 from both before and after the evaluated data segment, and all of the evaluated data in a segment is available for use in the calibration maps. This naturally results in more accurate vetoes, ranking, and significance.

In addition, the different stages of the analysis must be performed in sequence, that is the training is completed before the evaluation which is completed before the calibration and so on. The strides of data input to these different stages is longer than in the online case as well, mitigating the edge effects that occur when handling small data sets as discussed in the previous section. In reality, this mode is most frequently used for re-analysis of previous observing run data when the timescale of data availability is large.

The online configuration, however, uses SNAX for input feature vectors, and data is binned in the causal scheme described in section 6.2.1. This scheme allows for asynchronous production of trained models, evaluated datasets, calibration maps, and timeseries, but can only assign significance to data based on historical data. The models are trained every O(3) hours and keep information about historical data up to three days to inform the datasets, although this parameter is configurable. If there is a sharp change in the behavior of the detector, however, these models will mis-represent the current detector state. The data kept and used by the training processes to inform models are not weighted in any way by age, so keeping a long history of data for training makes the models and the calibration maps stable.

The calibration maps initialize the glitch and clean histograms with samples present in the training datasets from the latest model. Then, this is updated every few minutes with more recent samples from the evaluation processes which calculate ranks for new data in real time. If the look-back time for the training processes is decreased, then both the models and the histograms made by the calibration maps will update more rapidly with changing detector behavior, but when a new model is available and applied, the result could vary drastically from the previous one. This would make the output timeseries discontinuous which is itself undesirable. Therefore, a three day look-back is chosen as the balance between these two extremes for the models.

The timeseries processes then produce data in one second strides, continuously polling for the latest calibration map and model. This comes with issues discussed previously where information from vetoes is not kept across the one second strides, causing edge effects. Most frequently, the timeseries processes get a new model every few hours and a new calibration map shortly after. Thus, timeseries data is available at incredibly low latency, but could contain edge effects and is generated using models that are at worst a few hours old.

6.3 Support of Real-time Gravitational Wave Detection with iDQ

Ahead of the fourth observing run, iDQ was integrated into the live data quality reports (DQR) generated in real time for gravitational wave candidates, and the gravitational-wave analysis pipeline, PyCBC, integrated iDQ timeseries data into its real-time detection pipeline [162]. The DQRs cycle through a variety of data quality checks for each instrument in production as an

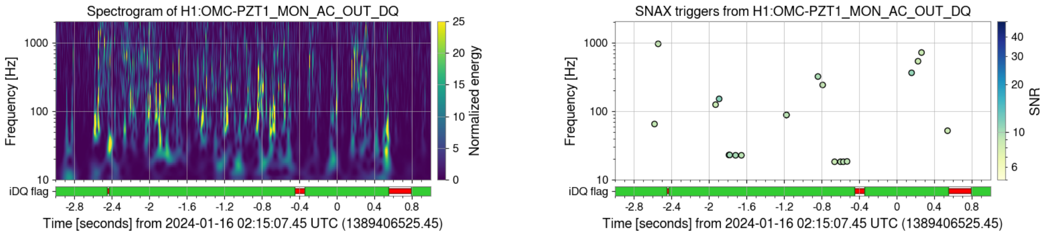


Figure 6.3. Example of iDQ follow-up task contained within the real-time data quality reports ahead of O4b. Left: A spectrogram of the auxiliary channel which ranked highest in model applied at the time. Right: The triggers from the auxiliary channel which ranked highest in the model applied at the time with the SNR assigned to them by SNAX. Below both figures is a vector accounting for when the veto corresponding to this channel was active.

immediate follow-up to a gravitational wave event. This includes checks on spectrograms of the strain data for glitch morphologies and on excess energy outside of the gravitational wave track, as well as initial checks on active auxiliary channels at the time. If the gravitational wave candidate fails any of these checks, a data quality expert is appointed to manually follow-up.

Most recently, iDQ was added to this list of checks. Now, iDQ information in a three second range around the time of interest is thresholded, and if the iDQ timeseries passes that threshold the check is failed. All previous data quality checks in the DQR only looked in the immediate vicinity, or about a second, around the time of the candidate. Adding this additional search from iDQ at a longer range enables data quality experts to estimate whether a glitch is present throughout the track of gravitational wave event. For example, in the case of GW170817, a glitch a few seconds before the merger time caused issues in detection pipelines, and this iDQ check could flag a similar situation.

Additionally, ahead of O4b, the iDQ report was improved to contain a spectrogram of the auxiliary channel which ranked highest in the model applied at the time of the event, and a corresponding plot of feature vectors SNAX generated from that channel as seen in Figure 6.3. While previously the auxiliary channel information was available via a list of the vetoes and their ranks, this additional step provides an immediate visual representation of that auxiliary channel allowing for verification of the auxiliary activity at a glance.

In support of both of these integrations, two additional changes were made to the online configuration of the iDQ analysis. A state vector was added which provides real-time information on whether or not the state of iDQ pipeline is production ready, and is used to alert downstream users of the quality of the iDQ timeseries. The state vector may indicate that iDQ is not production quality for a few reasons. In the first few hours of an analysis, there is not yet enough data collected in order to generate informed models or populate the histograms used by the calibration maps. Therefore, the number of samples in the training datasets must past a certain threshold for the state vector to reflect a production quality state. Additionally, models from the training data can become stale due to computing issues preventing the training processes from running. This can happen when there are issues in data distribution and discovery, or for a variety of reasons during maintenance on the computing center where iDQ runs. To mitigate this, we also apply a check on the age of the latest model, and if its stale then the state

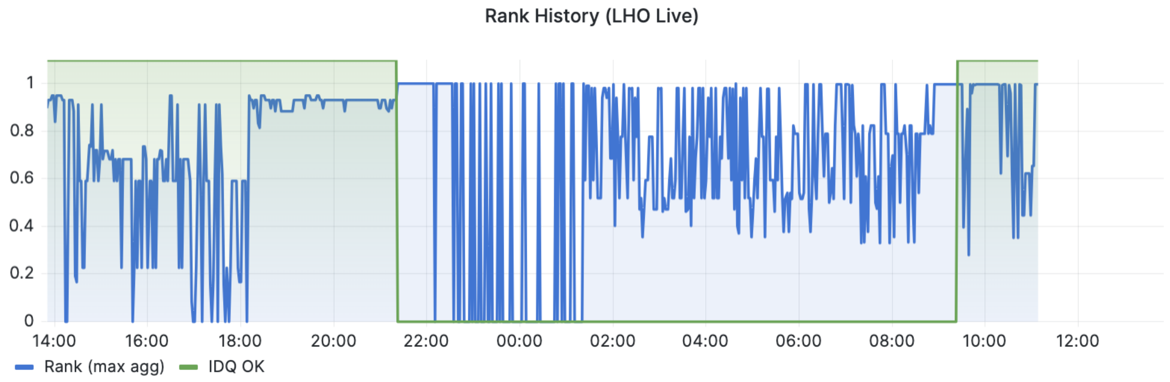


Figure 6.4. An example of one view available on the live monitoring dashboards implemented for each of the detector analysis during O4. It displays the live and historical timeseries output from the iDQ analysis along with the status of the IDQ_OK state vector.

vector is again set to reflect this.

With integration of the low-latency iDQ timeseries into downstream analyses, it also became prudent to have real-time monitoring of the entire analysis. We therefore integrated the use of Kafka, a low-latency event streaming platform, into the iDQ pipeline. This allows for diagnostic information such as the state vector, the output timeseries, and more to be passed asynchronously by any process to kafka, and stored in a database. The data is then pulled from this database and displayed with real time updates, and look-back capabilities via a Grafana dashboard.

Currently, all of the output timeseries from each of the monitored detectors is displayed with an overlay of the IDQ_OK state vector. This allows for a check on the state each analysis at a quick glance, and dynamic follow-up for any event time of interest that may appear in the low-latency analyses. An example of this can be seen in Figure 6.4.

Chapter 7
Performance of iDQ ahead of LIGO, Virgo, and KAGRA's Fourth Observing Run

7.1 Introduction

The detection of 90 gravitational wave candidates [43] by the LIGO Scientific Collaboration and the Virgo Collaboration has been made possible via gravitational wave detectors, Advanced LIGO [3] and Advanced Virgo [163]. The detectors are modified large-scale Michaelson interferometers whose main output, called strain data, is the differential difference in distance traveled by laser light between its two arms. As gravitational waves pass through the instrument, these paths change a minute amount, resulting in a recorded signal in the strain data measured to lengths as small as 10^{-19} meters. While this remarkable sensitivity allows for the detection of gravitational waves, it also allows for the easy detection of transient noise sources arising from the environment or the instrument itself. Such non-stationary noise sources are commonly referred to as glitches.

Glitches can arise from a variety of both known and unknown sources [24, 164–166] and pose a challenge to the accurate detection and analysis of gravitational wave signals [167–170]. Some glitches can manifest in the strain data in a similar manner to genuine gravitational wave signals, and can possibly be mistaken for them if present simultaneously across multiple detectors. Glitches can additionally overlap with true signals thereby obfuscating them, and making parameter estimation of the true signal difficult. Such was the case with the detection of the binary neutron star merger signal GW170817 where there was a large glitch overlapping with the Livingston detector data causing some initial concerns with the data quality [171, 172]. Therefore identifying and characterizing glitches is a key part of increasing the overall sensitivity of the entire detection system.

While the detection of gravitational waves is possible using the strain data alone, there are also thousands of supplementary data outputs produced by the detectors. These supplementary outputs, called auxiliary channels, record additional degrees of freedom in the detector outside of the strain data and act as monitors on everything from mirror deformation, to environmental recordings [166, 173–176]. Some of these auxiliary channels monitor systems which produce signals in the presence of gravitational waves. These channels are not ideal for use in identification of glitches as signals there could be genuine gravitational waves signals. Safe auxiliary channels then are defined those which are insensitive to gravitational waves, and therefore any signal present in these channels is by definition a glitch. Glitches witnessed by these safe auxiliary channels may also appear in the strain channel. If only monitoring the strain channel in such a case, information is lost that could easily identify the present signal as a glitch. iDQ is a statistical inference framework which uses safe auxiliary channels to identify these cases and make statistical statements about the presence of glitches in the strain channel based solely on activity in the auxiliary channels [11, 31].

iDQ is trained on activity in safe auxiliary channels labeled by the presence of glitches in the strain data to identify correlations between the two. If an auxiliary channel is identified to be strongly correlated with the strain data, then any new activity in that auxiliary channel can be used to predict if a signal in the strain data is of terrestrial origin. Additionally, by monitoring only the safe auxiliary channels, iDQ can safely identify glitches without also flagging real gravitational waves making it ideal for incorporation into gravitational wave detection pipelines. The output of the iDQ analysis then consists of two probability statements that indicate the likelihood that the gravitational wave data is contaminated by a glitch monitored by one of these auxiliary channels.

There are thousands of these safe auxiliary channels sampled at high rates available for analysis by iDQ. In order to reduce the computing cost and latency of analyzing so many channels at such high rates, iDQ relies on two sources for the extraction and downsampling of relevant information. In low-latency operation, the Stream-based Noise Acquisition and eXtraction pipeline (SNAX) [31], is implemented for this purpose, but in high latency Omicron [29, 30] is used and it is the latter which will be discussed in this work. Omicron reports on the presence of excess power, measured by the signal-to-noise ratio (SNR), in both strain and auxiliary channel data via the Q transform, a wavelet decomposition parameterized by a quality factor (Q). Times which are noted to have an estimated SNR greater than 5.5 are then passed on to iDQ for analysis. High SNR times from the strain channel are used to label glitches in iDQ training sets, while all times from auxiliary channels are used to find correlations.

Correlations identified by iDQ can then be further classified by the type of glitches present in the correlation. Glitches are separated into classes based on how they appear in the strain data and those which have the same morphology as gravitational wave signals, for example, are of particular interest and are targeted for further study. Although classification of glitches can be done by eye, the frequency at which glitches appear has motivated the development of GravitySpy [4, 25–27, 177] which can automatically classify any time of interest using a convolutional neural network (CNN). In this work, the correlations iDQ finds between auxiliary channels and strain data are extended using the classifications assigned by GravitySpy to find relationships between auxiliary channels and glitch classes. These relationships then reveal which glitch classes appear most frequently in which auxiliary channels. If an auxiliary channel frequently witnesses a particular glitch class, then the detector system the auxiliary channel monitors can be investigated as a possible source of that glitch class.

In this work, we first provide a background of the iDQ framework, the Omicron package it relies on, and the GravitySpy package which provides classifications. We additionally review changes to the iDQ framework from the LSC's third observing run to current day, as well as the performance of these changes. We then continue to quantify iDQ's current performance, including a new measure of glitch probability, and report on a new method using iDQ to track glitch types back to their possible origins in the detectors.

7.2 Background

7.2.1 Omicron

Omicron is designed to detect and characterize transient signals through the use of the Q transform, which decomposes the detectors' time-series data into a time-frequency basis. The Omicron implementation of the Q transformation relies on the tiling of a signal's time-frequency space where one tile is defined by the projection of the signal onto a Bisquare windowed sinusoid basis with a given central time, central frequency, and quality factor, Q [29]. The distribution of tiles used by Omicron in a single Q plane is defined by an acceptable energy loss due to mismatch between the tiles, and this strategy leads to sets of tiles defined logarithmically in central frequency and Q, and linearly in central time.

The excess energy of any given tile is then used as an estimation of the SNR, ρ and is given as [30]:

$$\rho^2 = \frac{|X(\tau, f_c, Q)|^2}{< X(Q)^2 >} - 2 \tag{7.1}$$

where $X(\tau, f_c, Q)$ is a single tile, $< X(Q)^2 >$ is the mean expected energy of all tiles in a given Q plane, and 2 is the result expected from white noise. In order to form events, any two tiles with positive excess energy which have less than 0.1 seconds between their central times are considered to be identifications of the same event and are clustered. After a cluster is formed, the SNR, central time, and central frequency of the tile in the cluster with the highest SNR are assigned as the event's parameters. Similarly, the start and stop time of the event are taken as the earliest and latest central time of tiles in the cluster [30]. This clustering results in events with non-uniform duration from as little as 0.1 seconds and up to 10 seconds, with the bulk of the distribution around at $\mathcal{O}(0.1)$ seconds.

Any event with an SNR of at least 5.5 in the strain channel is then recorded in the Omicron database along with its relevant parameters. High SNR events from the strain channel could indicate the presence of a glitch, but could just as easily identify a real gravitational wave signal. For the purpose of this paper, events with an SNR greater than 10 in the strain channel which do not correspond to known gravitational wave signals are treated as the identification of a glitch. Events identified in safe auxiliary channels, meanwhile, are used as inputs to the iDQ analysis.

7.2.2 GravitySpy

While other detector characterization pipelines like Omicron identify glitches in the strain data, GravitySpy contributes classifications to those times. There are variety of glitch classes defined by detector characterization experts based on their morphology in the time frequency space of the strain channel [25]. GravitySpy uses a combination of human volunteers and a CNN machine learning algorithm [4, 27, 28, 177] to classify any time of interest based on spectograms of the strain data with four different durations. Four durations are used in order to expose morphologies which are present at varying timescales. The output of GravitySpy then is a probability, or confidence, that a given time is of each glitch class considered. The confidence across classes is not required to sum to one, and instead each class is assigned a value between 0 and 1 individually. For example, GravitySpy could be confident that there was a scattering-like glitch present, but be unable to distinguish whether it was Fast Scattering or Scattered Light. In this case, you could expect high confidence values in those two classes and lower confidence values across all others.

The training set for GravitySpy has evolved with time as more glitches are discovered, and more morphological classes are defined. As of LIGO's third observing run, the training data consisted of 9631 labeled glitch samples across 23 morphologies [25]. There has since been an update to the GravitySpy model for LIGO's fourth observing run [27], but this paper focuses on data from the third and therefore uses classification from the GravitySpy model available during that time.

During active LIGO observing runs, GravitySpy classification is triggered on new Omicron event times uploaded to the database with SNR of 7.5 or greater. This allows for medium-latency identification and classification of glitches which can then be used by detector engineers to inform detector maintenance. However, Omicron experts frequently add new events to the database which were not identified in low-latency. These additional times are not always assigned classifications by GravitySpy, and therefore in this work, we consider an additional category of "Unclassified" to represent these times.

7.2.3 iDQ

Although well-described in [11], we will give a brief summary of the iDQ framework leading into the LSC's third observing run. iDQ runs in two modes – streaming and batch. We summarize the batch, or high-latency, implementation here as it is what is used to measure the performance of the pipeline in the later sections of this work.

iDQ begins by taking in events produced by Omicron on O(10^3) auxiliary channels. As discussed in section 7.2.1 triggers are equivalent to tabular data on transients in these channels and contain information on the signal-to-noise-ratio (SNR), frequency, central time, etc of these transients. The events for each channel reported by Omicron are then downsampled further by iDQ into feature vectors. A feature vector represents the maximum SNR event reported by Omicron in any one second window.

These vectors are then labeled as glitch or clean based on the SNR value reported by Omicron on the strain data at the same time. If the SNR of the strain channel is greater than 10, then the transient is considered a glitch. If the strain channel SNR is less than 5.5, then its considered clean. Any feature that falls in between those two thresholds is neither clean nor a glitch and is not used in training data.

To construct training datasets, the entire time of interest is first divided into segments. The majority of these time segments are used for training, and one is reserved for evaluation, as described in detail in Section IV of [11]. The times in training segments are then used to construct the training datasets iDQ needs for its classifiers. For training, all the times labeled as glitch, and a random selection of clean times at least one second away from a labeled glitch are used. The additional one second window for clean sampling enforces that the times in the glitch and clean datasets are uncorrelated as most glitches are shorter than one second in duration. This limits the training dataset based on the Omicron SNR threshold in strain, but it is ultimately the auxiliary channel features at the times of interest, and not strain information that iDQ is trained on.

iDQ uses these datasets to then train the classifier(s) chosen. The only limit to the choice of classifier is that it must map the high-dimensional input feature vector space into a single rank value between 0 and 1. This mapping, called a model, is unique to the classifier and to the time it was trained on. For batch production during O3, the classifier OVL was used.

The OVL classifier is well-described in [10], but, in short, models produced by OVL consist of an Ordered Veto List, a list of vetoes on the auxiliary channels which were active at the time of interest. To form this model, OVL creates a list of possible veto configurations based on auxiliary channel, threshold, and time window and then evaluates them based on a chosen metric. OVL previously supported the choice of one of three different metrics for this process: efficiency-deadtime ratio, Poisson significance, or use percentage, although as we describe in section 7.3.1, additional options have been added. The use percentage is the fraction of auxiliary channel glitches which can be associated with a strain channel glitch where a glitch in the auxiliary channel is defined by having an amplitude above the threshold set by the veto. The Poisson significance is the probability of observing as many or more coincidences between two series of random events than actually observed between the auxiliary channel and strain as described in detail in [178]. The efficiency-deadtime ratio is given by the efficiency of the veto over the deadtime introduced by the application of the veto. In other words, the fraction of total glitches in strain removed over the fractional livetime removed by applying the veto.

iDQ requires that these metrics produce ranks that fall in the space of [0,1) rank space in order to comparable between one another. If only used individually, the choice of [0,1) is an arbitrary one and what is more important is the ordering. However, as will be discussed in section 7.3, it can be desirable to combine information across several metrics in order to give preference to vetoes with a certain combination of properties. Therefore, a simple scaling factor was applied to their values with a map from the metric space of [0, inf) to rank space of [0, 1) as show below:

$$Rank_m = \frac{s_m * x_m}{1 + s_m * x_m} \tag{7.2}$$

where s_m is the scaling factor for the metric and x_m was the value of the metric itself.

After the initial rank evaluation, the vetoes are ordered from highest to lowest, and then the rank is re-calculated applying the highest veto first, and ending in the lowest. Any veto falling under a threshold for that metric is removed, and the process is repeated. In this way, OVL produces a final Ordered Vetoed List for any given training time, thereby making a model. The model for OVL is then applied to a time of interest by first removing any veto configuration which doesn't apply. Then, the rank of the highest ranked veto from the resultant list is applied as the rank of the time of interest.

The rank values alone, however, do not have any physical meaning and it is only the ordering that truly matters. Therefore, ranks must be transformed to log-likelihood and false alarm probabilities through calibration. To for this calibration, the rate of clean and glitch samples must be determined, and a PDF generated. Through O3, the glitch and clean distributions for this calibration were populated by the glitch and clean samples from the training datasets, although this has changed recently as described later in section 7.3.2. The rates of clean samples and glitch samples can then be calculated directly from these distributions as described in [11], and a Gaussian kernal-density-estimate (KDE) can be applied to the distributions to obtain a posterior-density-function (PDF). These rates, along with the PDF from the KDE, create the calibration maps needed to convert any given rank to a log-likelihood and false alarm probability as described in detail in [11].

In summary, for any time of interest, a model made by OVL and trained using Omicron labels on the strain is applied to the full feature vector set. The result is a single rank. The rank is then transformed to log-likelihood and false alarm probability statistics via a calibration map generated by sampling the clean and glitch distributions of the training datasets and applying a Gaussian KDE. This process can then be repeated for any number of times, thereby creating a full timeseries.

7.3 Pipeline Improvements

The methods described in 7.2 were applied uniformly through O3. However, between the end of O3 and the beginning of O4 in May 2023, several changes were implemented into the iDQ analysis to improve calibration, and the dynamic range of its outputs. These changes were applied to batch offline re-analysis of O3 data in order to prepare detection pipelines for the fourth observing run (O4) and then applied to both the batch and streaming iDQ analyses during the first half of O4. In the following three sections, we describe in detail the changes and the motivations behind them. In 7.3.4, we report on the performance of these changes and show that they lead to the desired effects.

7.3.1 Change of Rank

The first of these changes was to the calculation of the rank assigned to times of interest by OVL.

The metrics of use percentage, Poisson significance, and efficiency-deadtime ratio individually are useful, but previously it was difficult to compare results across them as they did not behave similarly across the rank space. A similar behavior of increasing metric leading to increasing rank with support throughout the [0, 1) rank range was needed for each metric. To get this behavior, the scaling factors described in Equation 7.2 for the efficiency-deadtime ratio and Poisson significance were altered but otherwise the rank is calculated identically. Meanwhile, the use percentage to rank map was changed entirely to have no scaling factor or map at all. That is, the rank is exactly the use percentage.

With this change, we then offer the option to combine ranks from multiple sources at once to obtain weighted ranks. This is done via a weighted average of the three individual metrics as shown below:

$$\frac{\sum r_m * w_m}{\sum w_m} \tag{7.3}$$

where w_m is a configurable parameter and is weight for a given metric, and r_m is the value of the rank from a given metric. In a typical offline analysis, the veto efficiency and use percentage of the veto are used with a weight of one third and two thirds respectively. This configuration down-ranks vetoes with high efficiency but poor use percentage. Vetoes which follow this trend flag auxiliary channels with rare departures above the threshold but in the process veto large periods of quiet time in the strain channel making the likelihood of false alarms increase. By prioritizing the use percentage metric, these veto configurations are suppressed in favor of those which are less likely to report false alarms.

7.3.2 Use of all data for background collection

In order to calibrate the rank from iDQ's classifiers into statistical information such as the false alarm probability and log-likelihood, a model of the underlying distribution of clean and glitch samples is generated as described in 7.2 using a Gaussian KDE. However, the KDE can only be as accurate as the underlying distribution it relies on.

As described previously, when using the binning and segment scheme developed for the offline batch mode of iDQ, the clean distribution used in calibration is populated only with times from the training datasets. This results in populating the clean distribution only with times at most one half of a segment's width away from the time of interest, or typically $O(3)$ days. When there are few segments for a wide time range, or the segment width is large, the clean distribution used for calibration can be very different from the true distribution which the time of interest resides in. This is a natural result of the time-evolving nature of the detectors as the noise background one day can vary significantly from the noise background the next.

This results in the output timeseries jumping between segment boundaries as the calibration between those segment boundaries reflects the change in the underlying noise distribution. By allowing sampling of the clean distribution within the time segment of the time of interest in addition to sampling outside of it, the resulting clean distribution more accurately reflects the local distribution, and the output timeseries becomes more seamless between segment boundaries.

7.3.3 Bounding of KDE Bandwidth

As mentioned in section 7.2, iDQ applies a Gaussian KDE to the discretely sampled clean and glitch histograms to create smooth posterior density functions. These distributions are then converted to cumulative distributions in order to calculate the false alarm probability (FAP) and log-likelihood. The accuracy of this KDE then has a direct impact on the timeseries output of iDQ.

Previously, the discrete nature of the clean and glitch distributions caused the automatic bandwidth optimization of the KDE to rail to extremely narrow Gaussians as shown in the top half of Figure 7.1. This left large regions of the rank space without proper support from the KDE, especially as rank goes to one. This caused scaling of the log-likelihood and FAP to

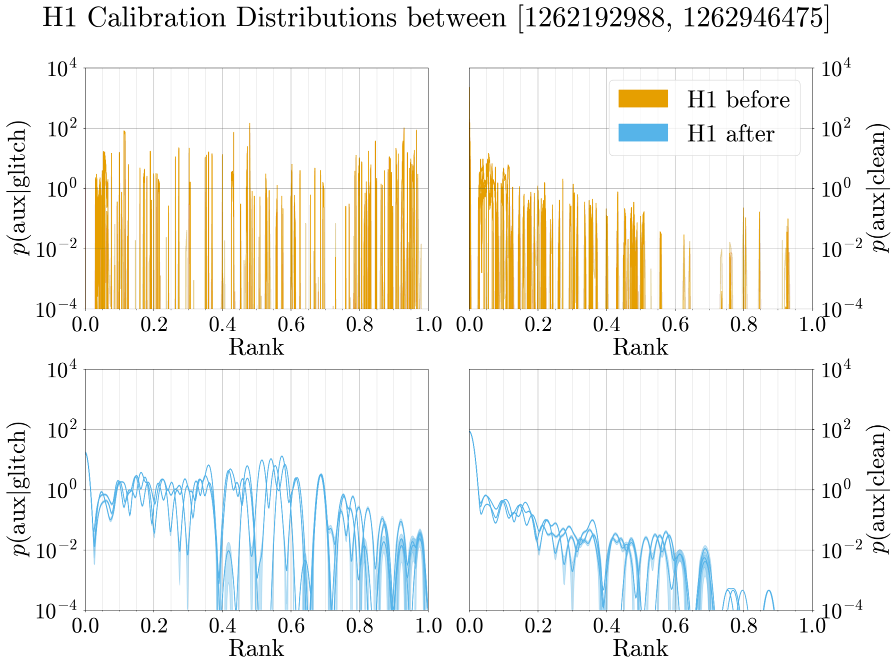


Figure 7.1. In the bottom row, the KDE of the glitch distribution (left) and clean distribution (right) of samples collected over about two weeks of O3b time after the most recent changes were implemented. In the top row, the cumulative distribution function calculated based on the corresponding KDE in the bottom row. Note how the KDE is significantly smoother than before and provides a wide range of support across the rank space. This results in the output statisitics based on this KDE being more evenly scaled across the entire rank space.

be uneven, and the log-likelihood in particular to have a large dynamic range. Bounding the lower end of the bandwidth range forces support in those portions of rank-space without many samples in its underlying histogram as seen in the bottom half of Figure 7.1. This results in a more evenly scaled output statistics, with a smaller dynamic range.

Improvements in the distribution and bounding of the timeseries can be seen in Figure 7.2. Here, we only show the results for Hanford because the Livingston timeseries show similar behavior changes. The top half of this plot shows the histogram of the log likelihood ratio between the glitch and clean models with the original timeseries shown in orange and the improved version in blue. The dynamic range of the log-likelihood ratio has been severely reduced from thirteen orders of magnitude to just six with the updates.

7.3.4 Performance of Pipeline Changes

In order to show the improvements made in the pipeline, we compared results from iDQ's analysis of O3b, which spans from November 2019 to March 2020. We compare the results from the original offline analysis done concurrently with O3b observations and from a re-analysis using the newly updated pipeline code. For simplicity we will refer to the different code versions as being 'before' and 'after' respectively.

Figure 7.3, shows the false alarm probability (FAP) plotted against the efficiency of the

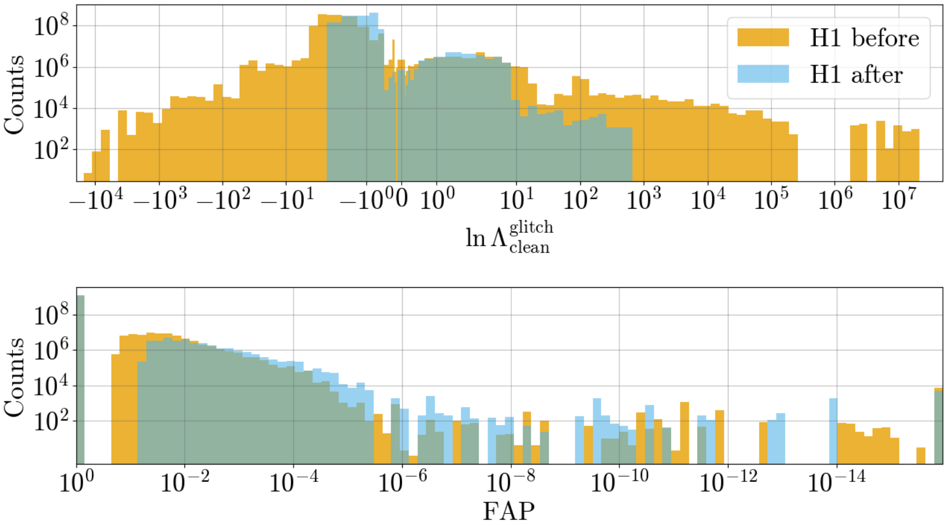


Figure 7.2. Histograms showing the count of samples with log likelihood ratio between the glitch and clean models (top) and false alarm probability (bottom) observed by iDQ at Hanford during O3b both before code changes were implemented (yellow) and after (blue). Notably, there is a vast improvement in the dynamic range of the log-likelihood distribution after the code changes as desired.

iDQ pipeline at Hanford (red) and Livingston (blue) with original code (solid) and the updated version (dashed). In these receiver operating characteristic (ROC) curves, we can already see a stark contrast between the two code versions – particularly in the performance at the Hanford detector. Both at Hanford and Livingston, there is a doubling of the efficiency of the pipeline at a false alarm probability (FAP) of 10^{-3} after the updates and a general improvement across the range of FAP 10^{-3} to 10^{-1}. At larger FAP, the model for Hanford previously did worse than an uninformed one, or a model built on random chance, but after the updates this is no longer the case. However, the Livingston performance is slightly worse at high FAP than before. Generally, this slight decrease in performance at high FAP is seen as a more than fair trade-off for the wide improvement in the middle range. At high FAP, iDQ loses its distinguishing power as there is a gap in the rank output of OVL between 0, and the lowest ranked veto configuration as can be seen in Figure 7.2. This means that iDQ already does not have distinguishing power in the FAP range, so a small loss of sensitivity there is not a large loss to the power of the analysis. Instead, its the middle ranges of the FAP where we see the most improvements which are the most crucial to the distinguishing power of the analysis.

7.4 Use of iDQ

iDQ's main purpose is to identify glitches apparent in the strain data by monitoring the auxiliary channels of the detectors. Omicron strives for the same end goal, but via excess power in the strain channel itself. In this section, we use the glitch times flagged by Omicron as a benchmark for comparison for the output of iDQ and then analyze the results broken down by glitch class and auxiliary witness channel.

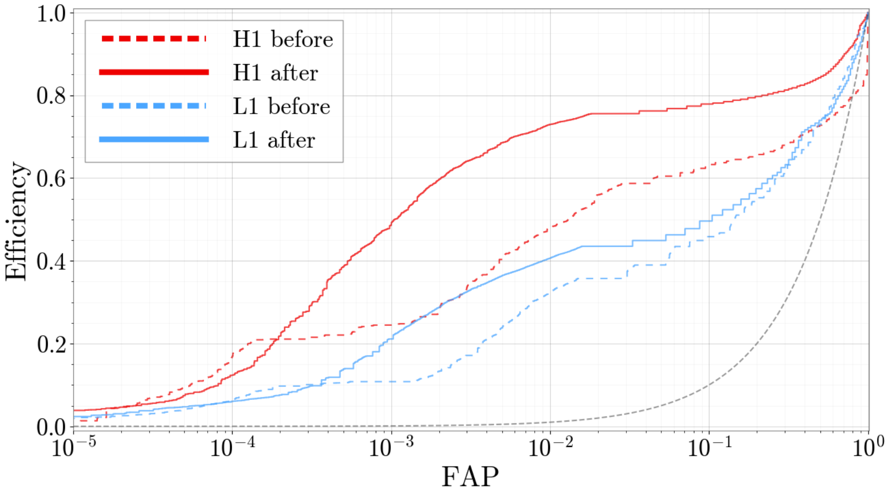


Figure 7.3. Receiver operating characteristic curve of iDQ at the Hanford (red) and Livingston (blue) gravitational wave detectors over the time period of O3b. The dashed lines represent the original results from iDQ and the solid lines show the re-analysed results after code improvements. A dashed grey line was added to represent an uniformed classifier. The analyses at both detectors show improvement in the important mid-FAP range where the bulk of iDQ's distinguishing power.

In Section A, we discuss how Omicron and iDQ identify glitches, how we construct coincident events between the analyses, and how the results break down by glitch class. In Section B, we show how iDQ can provide additional auxiliary channel information about identified glitches, and possibly identify physcial sources of glitch classes.

7.4.1 Glitch Presence Identification

7.4.1.1 Methods

In this work, we compare iDQ's performance against that of Omicron broken down by glitch type as classified by GravitySpy. In order to define a glitch as identified by iDQ, we first apply a threshold on the log-likelihood ratio. In this work, we've chosen two thresholds to examine – two and five. We have chosen to analyze both of these because we found certain glitch classes ring up frequently in the 2 to 5 range, but the lower confidence threshold additionally results in more false alarms. The nature of veto application causes the output timeseries of iDQ to be step-wise with steps at most the width of the largest veto window, but typically less than a second. Therefore, after application of the threshold, we cluster any points with identical adjacent neighbors by keeping only the central point in time from series of identical points. This is equivalent to identifying the center of any veto window as the central time of the event. We then further cluster these points by keeping the maximum log-likelihood point in a clustering window, w_{iDQ}, of one second, or a half second on either side. Using this large clustering window allows us to assume that any two event times identified by iDQ are not caused by the same glitch, and are uncorrelated.

We then compare these events identified by iDQ to all times identified by Omicron in the

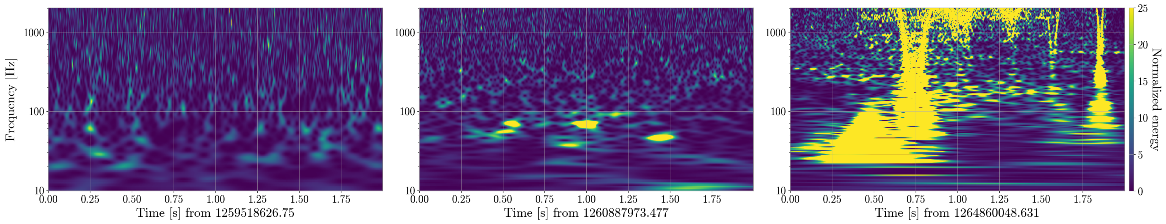


Figure 7.4. Qscans of three times identified by both Omicron, and iDQ as being a possible glitch, but which was classified by GravitySpy as the category "No Glitch". The left-most plot shows a correct classification of a No Glitch time while the middle and right show incorrectly classified times. The leftmost figure shows excess power, but fairly evenly distributed with no clear concentration – a true No Glitch classification. The middle shows a clear time of excess power in a distinct shape, but not in a morphology that matches any other glitch class. The right most plot shows a time which is clearly an Extremely Loud glitch, but was just mis-classified by the model.

strain channel as having an SNR of ten or greater. The threshold of ten on Omicron SNR is chosen to match the SNR threshold used in iDQ training datasets, and we calculate the Omicron glitch rate, σ_{omic} as the number of events crossing this threshold by the observing time. Any iDQ event within a coincidence window, w_{coinc}, of a half second on either side of a glitch identified by Omicron we assume is an identification of the same glitch and call this event coincident between the analyses. We can then count the number of these events for the entire observing time, and call that number N_{coinc}.

To classify these coincident glitch events, we find the classification reported by GravitySpy for the Omicron event in the strain channel and enforce that the confidence of the classification is greater than 0.9. We then assume that this classification applies to the Omicron event, and the relevant coincident iDQ event as well. In addition to known morphological glitch classes, GravitySpy additionally includes one classification called "No Glitch". This class is meant to truly classify times without glitches present, but the GravitySpy model used in this work has recently been found to confidently assign this label to times which clearly have excess power in their spectrograms that does not necessarily match any of the other classes. Three examples of this classification can be seen in Figure 7.4. The leftmost panel shows a correct classification of No Glitch. While there may be some excess power, it is not noticeable above the noise background nor well-localized and is therefore not a glitch. The middle panel shows a time which is a glitch, but which does not match the morphology of any glitch class known to GravitySpy, and is instead mis-classified into the No Glitch category. In the far right panel, we show a clear Extremely Loud glitch that mis-classified into the No Glitch class. It's possible that the overlap of the Extremely Loud glitch with loud repeating whistle glitches confused the classifier, but it is certainly not devoid of a glitch. In order to avoid confusion, we therefore throughout this paper will reference this "No Glitch" category as simply "Unknown" as these times may or may not contain glitches.

Any iDQ event which crosses the log-likelihood chosen, but which does not coincide with an Omicron event with SNR greater than 10, we assume to be identification of a time which does not contain a glitch and is therefore a false alarm. The iDQ training sets meanwhile use times with an Omicron threshold less than 5.5 or no Omicron event at all, for identification of times which do not contain glitches. Times which fall in the range of Omicron SNR 5.5 and 10 are not confidently true glitches, but do contain excess power in their spectograms as identified by Omicron. Thereby assigning false alarms in this work to be any event not coincident with an Omicron event with SNR greater than ten is a conservative estimate. We then calculate a false alarm rate as the number of false alarms in a given time period over the total detector observing time during that period.

We then additionally calculate the rate at which these coincidences would appear for two Poisson event generators. If the rate at which coincidences actually appear is greater than the Poisson rate, then we can conclude a true correlation between the events which iDQ and Omicron report with a estimate of the significance as the ratio of the coincident rate to Poisson rate. We calculate this Poisson rate as follows:

$$\sigma_P(\mathcal{L}) = \sigma_{omic} * \sigma_{iDQ}(\mathcal{L}) * T \tag{7.4}$$

where σ_{omic} is the rate of omicron events with SNR greater than 10, $\sigma_{iDQ}(\mathcal{L})$ is the rate of iDQ events above the chosen log-likelihood threshold, $\mathcal{L}$, and T is the total observing time.

We further construct a probability on the data being a glitch given what we have observed using Bayesian statistics:

$$P(\text{glitch}|\text{data}) = \frac{P(\text{glitch})P(\text{data}|\text{glitch})}{P(\text{data})} \tag{7.5}$$

Where P(glitch) is the prior and P(data) the normalization. P(glitch) can be taken as the probability of observing a glitch independent of iDQ and in this study is the probability of any time being flagged by Omicron with an SNR greater than 10. Assuming glitches are Poisson distributed, and that Omicron is effective at identifying them, we can calculate the probability of observing at least one glitch per coincidence window as:

$$P(\text{glitch}) = (1 - \exp(\sigma_{omic} * w_{coinc})) \tag{7.6}$$

where σ_{omic} is the omicron glitch rate rate, and w_{coinc} is the half second coincidence window as described before. P(data) is then the probability of seeing iDQ data above the threshold we chose, or the probability that the time of interest has been flagged by iDQ. We can calculate this probability similarly:

$$P(\text{data}) = (1 - \exp(\sigma_{idq}(\mathcal{L}) * w_{idq})) \tag{7.7}$$

where again $\sigma_{idq}(\mathcal{L})$ is the rate of iDQ events above the rank threshold, and w_{idq} is the clustering window as described previously. Finally, P(data|glitch) is then the probability of having seen the iDQ data given that a glitch is present, or the probability of the iDQ data occurring given that there is also an Omicron glitch flagged. In our study, this must be dependent on the ratio of total time covered by coincident events, and the total time covered by glitch events. In other words:

$$P(\text{data}|\text{glitch}) = (1 - \exp(\eta_{coinc}(\mathcal{L}))) \tag{7.8}$$

$$\text{where } \eta_{coinc}(\mathcal{L}) = \frac{\sigma_{coinc}(\mathcal{L}) * w_{coinc}}{\sigma_{omic} * w_{coinc}} = \frac{\sigma_{coinc}(\mathcal{L})}{\sigma_{omic}} \tag{7.9}$$

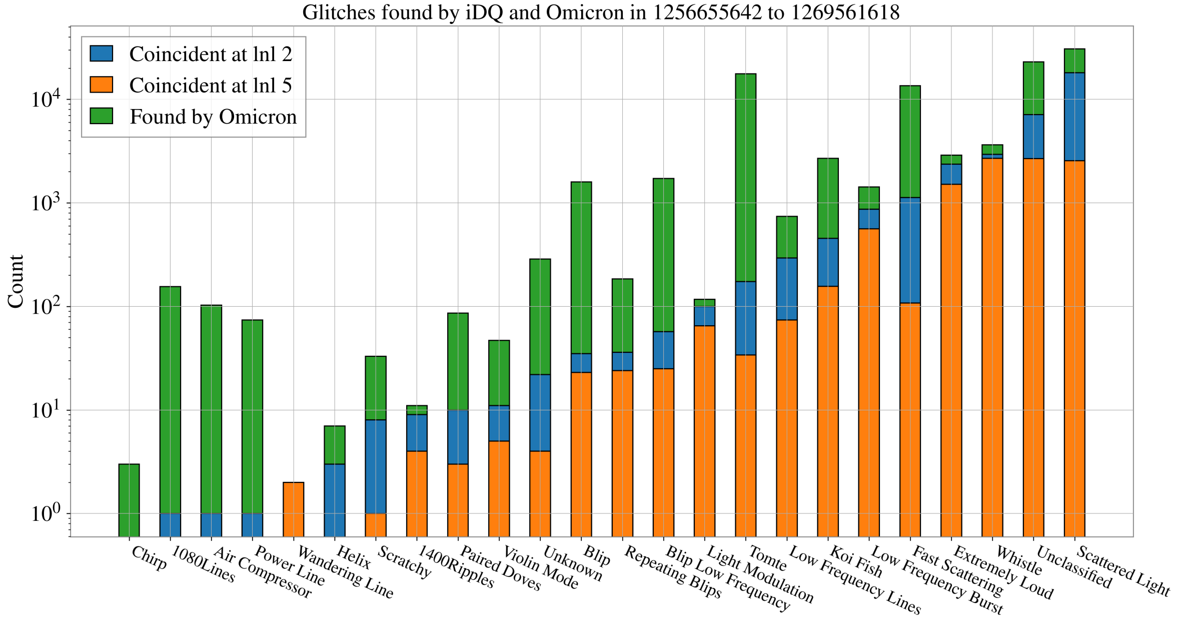


Figure 7.5. Glitch classification of event times in O3b. In blue and orange, the coincident times between Omicron triggers with snr greater than 10, and the times which pass two different thresholds on iDQ log-likelihood. In green, the times in strain data where Omicron reported SNR greater than 10, but not passing the log-likelihod threshold in iDQ. As shown by the orange and blue bars closely matching the green, there are several glitch classifications which iDQ seems particularly good at identifying including Scattered Light, Whistle, Extremely Loud, and Low Frequency Burst. This means that during O3b, iDQ likely had extremely effective witnesses for these glitch types, while there may not have been good auxiliary witness channels for others, like Tomte glitches.

The final probability can then be constructed as:

$$P(\text{glitch}|\text{data}) = \frac{(1 - \exp(\sigma_{omic} * w_{coinc})) * (1 - \exp(\eta_{coinc}(\mathcal{L})))}{(1 - \exp(\sigma_{idq}(\mathcal{L}) * w_{idq}))} \tag{7.10}$$

This estimate of the probability has the benefit of being based solely on counting statistics, meaning the underlying distributions can be collected cumulatively for real-time analysis without loss of latency. This estimate has the downside, however, of being dependent on the coincidence, and clustering windows chosen to estimate the duration of glitches identified by iDQ and Omicron. As constructed, the maximum $P(\text{glitch}|\text{data})$ obtainable is dependent on the ratio of the coincident and idq clustering windows. For example, implementing a coincident window half the size of the idq clustering window results in a maximum obtainable $P(\text{glitch}|\text{data})$ of one half. To mitigate this, the constant windows could instead be replaced by time-based segments flagged by iDQ constructed with the un-clustered iDQ timeseries, and by Omicron constructed with the individual glitch durations. Then, the time covered by coincident events could be given as the overlap between the two sets. The implementation for low-latency, and this definition using segment logic, has been left to future work.

7.4.1.2 Results

We take our results over all of O3b, or November 2019 to March 2020 from the LIGO Livingston detector. During this time, there were 100,512 number of departures identified by Omicron with a SNR greater than 10 available in the Omicron database. 78,436 of these additionally had available classifications by GravitySpy with a confidence greater than 0.9. Using the clustering and coincidence methods described in section 7.4.1.1, iDQ identified 39,398 (39 percent) at a log-likelihood threshold of 2, and 11,915 (12 percent) at a threshold of 5. It is then evident that iDQ identifies only a fraction of the total number glitches identified by Omicron, but this is expected. iDQ can only identify glitches for which there are auxiliary channels that reliably predict their presence. If there is no auxiliary channel activity in any of iDQ's witness channels at the time of a glitch, then iDQ can never report on it. This can be the case for glitch types whose source channels are not currently monitored by an auxiliary channel in the witness list, whose source may only register quietly in the currently monitored channels, or whose source is not monitored at all by any current auxiliary channel.

The classifications of these coincident events can be seen in Figure 7.5 in blue and orange while all of the Omicron events are shown in green. As mentioned previously, the Unclassified category comes from coincident events which did not have a classification from GravitySpy with a confidence of more than 0.9. During O3b, it is clear from Figure 7.5 that Scattered Light glitches were the main category of glitches plaguing the detectors with over 30,000 present while Tomte and Fast Scattering glitches are closely tied for second. Additionally, this figure shows that iDQ identifies a large fraction of some of the most common glitch types like Scattered Light, Whistle, Extremely Loud, and Low Frequency Burst while it struggles to identify others like Tomte, Fast Scattering and Koi Fish. As previously mentioned, this likely means that there are auxiliary channels which reliably record the presence of Scattered Light, and Whistle glitches while there may have been a lack of such channels for Tomtes and Blips.

Additionally, the fact that iDQ does not report on a single Chirp glitch at either threshold is, in fact, a feature. Chirps, as labeled here, are times in the strain channel which have a high SNR and whose morphology mimics a chirp shape. In other words, these are times which could very well be real gravitational waves as real gravitational wave events also follow the chirp morphology in strain data. As a data quality product, it is desirable that iDQ does not identify these times. For example, if the output of iDQ were used to generate vetoes for search pipelines, we would not want times which could be gravitational waves to be included in the vetoed set. This is an advantage of using software like iDQ which depends only on the auxiliary channels of the detector, and which therefore is insensitive to gravitational waves versus software like Omicron which directly analyzes the strain channel and identifies both real gravitational waves and glitches without delineation.

In Figure 7.6, we show the false alarm (blue star), Poisson (green circle), and coincident glitch (orange triangle) rates at a log-likelihood threshold of two (top) and five (bottom) as a function of time during the course of O3b. Although this data covers all of O3b, it was broken into shorter chunks for evaluation, defined by convention within the LIGO Collaboration. Each chunk of data corresponds to about two weeks of coincident observing time, and three bins were in the offline analysis resulting in evaluation segments, or each x tick, being between 2-4 days apart. Every three points then correspond to the same chunk of data, and the start

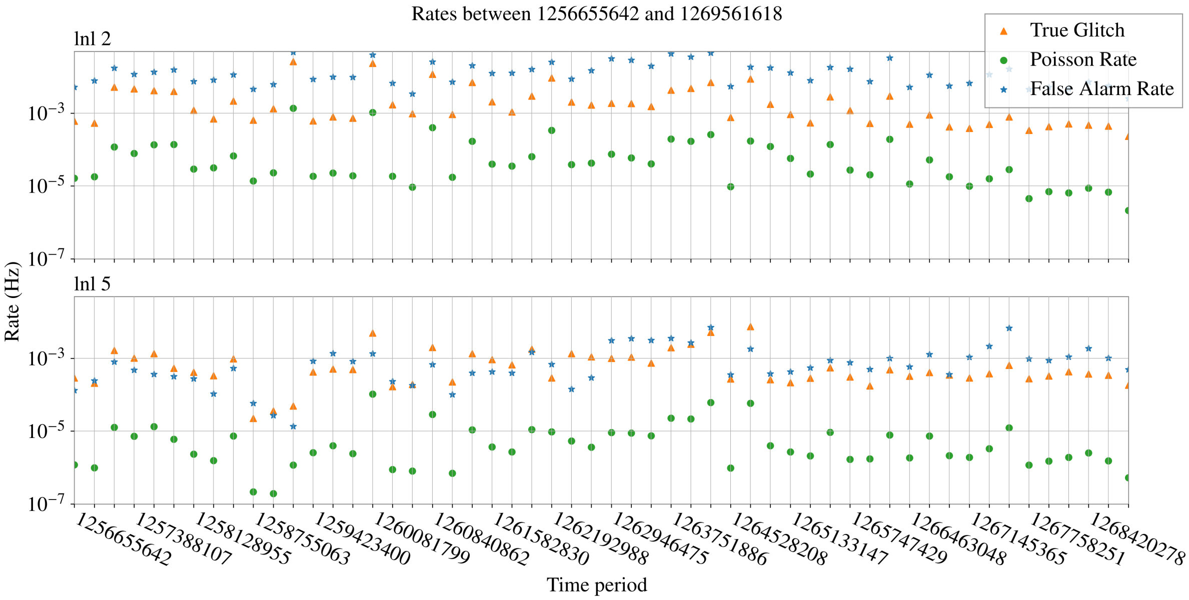


Figure 7.6. True glitch (orange triangle), false alarm (blue star), and Poisson (green circle) rates as reported by iDQ at a log-likelihood threshold of two (top) and five (bottom) over the course of O3b broken up by approximately five day periods. A true glitch is considered to be an iDQ time crossing the threshold which is coincident with an SNR greater than 10 Omicron time while a false positive is one not coincident with such an event. More details on this delineation are discussion in 7.4.1. Notably across time periods, the true glitch, or coincidence, rate is always at least an order of magnitude more than the poission rate implying a true correlation between iDQ and Omicron triggers. Additionally, the false alarm rate is about the same as, or higher than the glitch rate at the log-likelihood threshold of 2. However, at a threshold of 5, the this relationship begins to switch for some time periods.

of a new chunk is delineated with a GPS time label. At both thresholds, the coincident glitch rate is always at least one order of magnitude larger than that of the Poisson rate. This shows that the coincidences formed between iDQ and Omicron are more significant than random chance, and are truly correlated. As the log-likelihood threshold for iDQ is increased, both the glitch and false alarm rates decrease - a natural result of the increasing confidence a higher log-likelihood corresponds to. Additionally, as the threshold increases, the true glitch rate is more frequently higher than the false alarm rate than at the lower threshold, again showing the increase in confidence.

At the higher log-likelihood threshold, the variation over time is especially notable. The difference in time between points is only a couple days, and there are occasional jumps in rate more than an order of magnitude between neighboring points. This highlights the occasionally extreme variation in the noise background of the detectors even over the course of just a few days and the challenges that detector characterization experts face in characterizing this behavior. Additionally, the latter half of this plot reveals an interesting behavior change in both the detectors and iDQ. The glitch rate peaks at both thresholds just before 1264528208, or February 2020. Just after, however, the variation in rate at the detectors settles significantly and at both thresholds the false alarm rate is always higher than the coincident rate. This could point to some notable change in the underlying behavior of the detectors around that time which is propagating into the effectiveness of the auxiliary channels which iDQ uses.

In Figure 7.7, we show the results of Equation 7.10 using data from the entire course of O3b on the left and over about a two week period on the right. The x-axis in both plots represents the varying threshold applied to the iDQ event times used in the calculation of coincidence and data rates with the combined rank shown in equation 7.3 as the metric. The plot shows the steadily increasing probability of a glitch time with increasing rank up to a rank $\tilde{0}.8$, or log-likelihood of about 200. This is expected as the increasing log-likelihood corresponds to an increasing confidence by iDQ of a glitch presence. The plateau at the smallest ranks between 0 and about 0.1 are because events with log-likelihood values less than two, or about a rank of 0.1, were not included in this analysis and therefore we see a plateau extending to 0 at the logl 2 value.

After rank 0.8, there is additionally a dip for higher threshold times where we would expect a continuation of the upward trend. This is because the current implementation of the combined rank with a 2/3 weight on use percentage favors vetoes which activate as few as a single time during the training period which happens to coincide with a glitch time. This gives the veto a use percentage of 100% as it correctly flagged a glitch during the single time it was active, and is therefore highly ranked. However, vetoes such as these do not generally predict glitch presence, and therefore cause false alarms when applied to real data, causing the turn we see at high rank.

To mitigate this, we enforce that vetoes must also have a minimum Poisson significance to be considered. The image on the right of Figure 7.7 shows the results for a single chunk, or about two weeks, out of a total 17 chunks of data from O3b with a variety of Poisson minima enforced. At the smallest minmum Poisson significance value, we see the highest rank P(glitch) values go to what we would expect, but with those at slightly lower rank still being affected. As you increase the minimum Poisson significance, however, we see the effect extend to the lower rank values as well until eventually at a minimum of 20, the slope extends fully across all ranks. While the minimum Poisson significance has not been implemented across the O3b data in this work, this change will not affect the results shown here as the thresholds applied were at low log-likelihood values which live at an equivalent rank much smaller than 0.8.

Compared to the prior probability given by Omicron of 3% for this same data, the P(glitch) values across the rank space are an improvement over using Omicron information alone. At the lowest threshold considered in this paper of a log-likelihood of 2, or equivalently rank of around 0.1, P(glitch) already sits at a value of 14%. This shows the power in combining results from across both data quality products.

7.4.2 Auxiliary Channel Witness Identification

An additional benefit of an iDQ identification is that we can glean further insight into these glitch types through the auxiliary channels which monitor them. While GravitySpy allows us to classify times identified by iDQ and Omicron, we further this classification by combining the GravitySpy label with feature information on auxiliary channels used by iDQ. Each time analyzed by iDQ is assigned a rank via an Ordered Veto List by OVL as described in section 7.2. In this analysis, we take the model applied at a time of interest and then look at all of the vetoes which pass the log-likelihood threshold in the relevant Ordered Veto List to discover

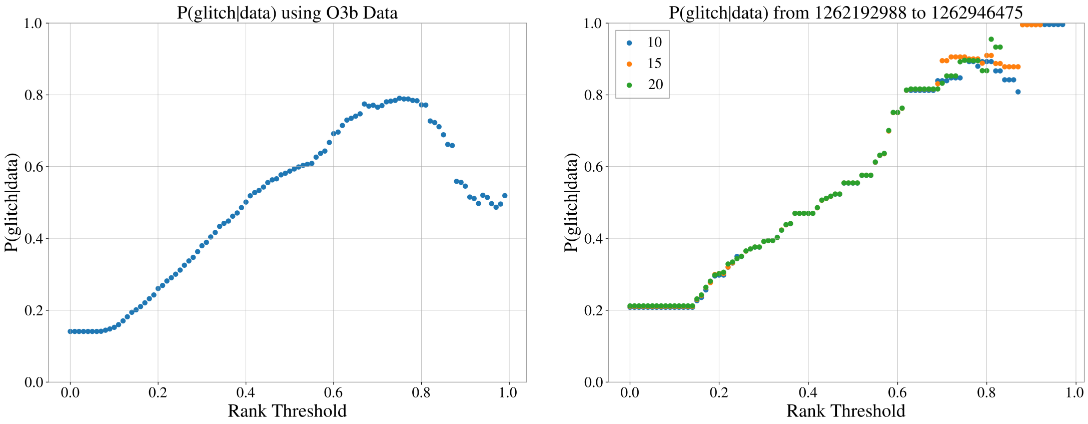


Figure 7.7. P(glitch|data) as defined in equation 7.10 using coincident events over the entire course of O3b (left) and one approximately two week period (right) between Omicron with SNR > 10 and iDQ events at a variety of thresholds on rank. On the left, this plot demonstrates the steadily increasing probability of a time being a glitch with increasing iDQ rank up to a certain point around a rank of 0.8. After this, the implementation of use-percentage in the combined rank value causes a number of vetoes with very small Poisson significance to be ranked highly, thereby causing a sharp increase in false alarms at high rank. On the right, a demonstration of how enforcing a minimum Poisson significance on vetoes mitigates this behavior. Already at a threshold of 10 on the Poisson significance, the turning point is mitigated. Then at a threshold of 20, the turning is completely removed.

which correlated auxiliary channels were active.

The channels associated with the vetoes which pass the log-likelihood threshold are then by definition those which make the most effective vetoes, or those whose activity often corresponds to excess power in the strain channel. It is not unreasonable then to assume that the channels which made the best vetoes for a certain glitch class at the very least frequently record instances of it, and at best could monitor the subsystem which is a possible source of that glitch class. Under this assumption, we look across all times of interest and count how often individual channels are active during a particular class of glitch.

In Figure 7.8, we show this correlation using a log-likelihood threshold of two and weighting by the total number of glitches recorded by Omicron. In Figure 7.9, we show this same correlation at the same threshold, but weighted only by the total number of glitches recovered in coincidence by iDQ and Omicron at that threshold. The former gives an overview of channel performance relative to all glitches of that class and identifies channels which are particularly good witnesses of most glitches in that class. It is possible, however, that there are multiple sources for the same class of glitch, so the latter plot focuses only on the sources which iDQ identifies instead of on the general performance of these channels. Additionally, we choose the lower log-likelihood threshold of two in this case to use as much information available to us as possible, and include information on clean samples and false alarms as a protection against the lower confidence.

On the y-axis of both plots are the channels which are active at least 50% of the time for at least one glitch class. On the x-axis are the glitch classes with at least ten coincident events, and which have at least one channel that appears at least 20% of the time for at least one glitch class. In Figure 7.8 we additionally include one column which reports on the auxiliary channel presence during the false alarms recorded by iDQ and weighted by the total time not covered by coincident glitches e.g. clean time, as there is by definition not coincident Omicron triggers available for False Alarms. This is a stand-in for the probability that the channel was active during a random non-glitch time.

In Figure 7.9 we again include an additional column of auxiliary channel presence during false alarms, but weighted by the total number of false alarms across all channels. In both plots, the Clean and False Alarm columns are to exclude the possibility that a channel seems to be a good witness for glitches simply because it is almost always active. As described in Section 7.2, channels which are active the majority of the time generally get down-ranked by the OVL ranking scheme because the high activity introduces a large deadtime and low use-percentage to the veto. Therefore, we wouldn't expect these kinds of channels to appear highly ranked in the veto lists and in these plots, but we consider all vetoes passing the log-likelihood threshold in the OVL list without weighting them. Therefore channels could be frequently ranked middling to low across glitch types, but appear in these plots to have the same significance as one which constantly appears with the highest rank thereby inflating its apparent significance. The inclusion of auxiliary witnesses for clean and false alarm samples is a sanity check on this possibility. If a channel is frequently present across glitch types, and during clean or false alarm times then we know that its possible this channel is just generally active. However, if it is present during glitch times and not during clean or false alarm ones then we can be confident that it a true witness for glitches.

The auxiliary channel ASC-X_TR_A_NSUM_OUT_DQ, for example, seems to be one of these extremely active channels. In both Figure 7.8 and 7.9, this channel appears incredibly frequently across glitch types. In Fig 7.9, however, it appears at over 70% during the False Alarm events as well. This is the perfect example of one of these generally active channels which results in a veto with a high efficiency, but middling use percentage. In other words, it creates an efficient veto which flags many glitch classes, but also can flag false alarms and we're seeing that trade-off here. This kind of veto may appear at middling to low rank in the ordered veto lists, but still cross the log-likelihood threshold of two and therefore appear in this analysis.

The three channels below that one in Figure 7.8, however, are good examples of the opposite. ASC-X_TR_A_PIT_OUT_DQ, ASC-X_TR_B_NSUM_OUT_DQ, and ASC-X_TR_B_PIT_OUT_DQ each appear over 50% of the time for all Scattered Light glitches, or more than 90% of the time for all coincident Scattered Light glitches over the course of O3b in Fig 7.9 while appearing significantly less frequently for all other glitch types including only a little more than a third of the time for False Alarms. This implies that while these auxiliary channels may be excellent witnesses of one source of Scattered Light glitches, and therefore are present in 90% of the ones flagged by iDQ, there are likely other sources which do not have good auxiliary witnesses contributing to the overall number flagged by Omicron. These three channels specifically monitor different pieces of the same subsystem in the Livingston detector. They are each part of the angular sensing and control subsystem (ASC) in the direction of the X arm, particularly monitoring the transmitted light (TR) on two different photodiodes (A/B). It follows that channels detecting transmitted light on one of the major axes of the detector would observe a large fraction of scattered light from the high powered lasers passing through the main mirrors, and indeed Scattered Light glitches are known to be caused by this system. Its interesting though that while one channel in this susbsystem set, ASC-X_TR_A_NSUM_OUT_DQ, is active across glitch classes and false alarms, three others catch more exclusively true Scattered Light glitches and this demonstrates the power of this method. From this information, its safe to say that X arm beam in particular at Livingston during O3b was a source of one of the most common glitches plaguing the detector at the time, particularly the three origins of the specific channels mentioned above. In fact, this is known to be the case as discussed in detail [24]. Halfway through O3, ground motion was found to cause variation in the differential difference between mirrors in the arms, and therefore cause scattered-light to rejoin the main beam registering as Scattered Light glitches. During the commissioning break between O3a and O3b, maintenance was performed to mitigate this issue which drastically decreased the total number of Scattered Light glitches recorded during O3b comparatively to O3a, but as is clear from this data, some glitches persisted from this source.

ASC-REFL_A_RF9_Q_PIT_OUT_DQ is another excellent witness, particularly of Whistle glitches. This channel appears in over 70% of all Whistle glitches recovered during O3b, and in almost 95% of the ones recovered in coincidence while only appearing in 11% of false alarms recorded by iDQ. It's sister channel ASC-REFL_A_RF9_Q_YAW_OUT_DQ is also highly active at over 55% of all Whistle glitches and 60% of the coincident ones. Both channels again monitor the angular sensing and control grouping, but particularly monitor the reflected light (REFL) of the power recycling cavity landing on a particular photodiode (A) in the pitch (PIT) and yaw (YAW) directions. It's interesting that these Whistle glitches are be observed so well by a channel monitoring reflected light, as reflected light is more often attributed to various kinds of scattering glitches. It's possible that this channel just happens to be downstream of the system which is actually generating the glitches, and so it is being monitored here despite not being the origin. However, the very similar monitoring channels on photodiode B show extremely similar correlations to the Whistle glitch types as these two channels, so it becomes even more convincing that a Whistle glitch source lies either in or upstream of these reflected light monitors.

These four channels are additionally active for a handful of other glitch classes as well such as the Blip and Repeating Blip class, although with a weaker correlation. Blip glitches could have the same sources as Repeating Blips as arguably one is just a more frequent version of the other. It would be new information though to also include Whistle glitches in that mix. It is possible that this could be a false correlation if the GravitySpy model, for example, frequently misclassified Whistle glitches as Blips and vice versa, but there is no evidence that that is the case. Instead, this could be the result of a couple different configurations. This channel could monitor the physical sources of multiple glitch classes, meaning it could be downstream from other parts of the detector system which individually generate these glitch types. It could also be that there is one source which this channel monitors that creates many different kinds of glitches. Generally, we assume that glitch classes are generated by distinct sources, but if this latter situation is true it could hint at correlations between glitch classes previously unconsidered. With this information alone, it is hard to tell which may be the true case, but either way it demonstrates the usefulness of this method in further characterizing detector behavior.

7.5 Conclusion

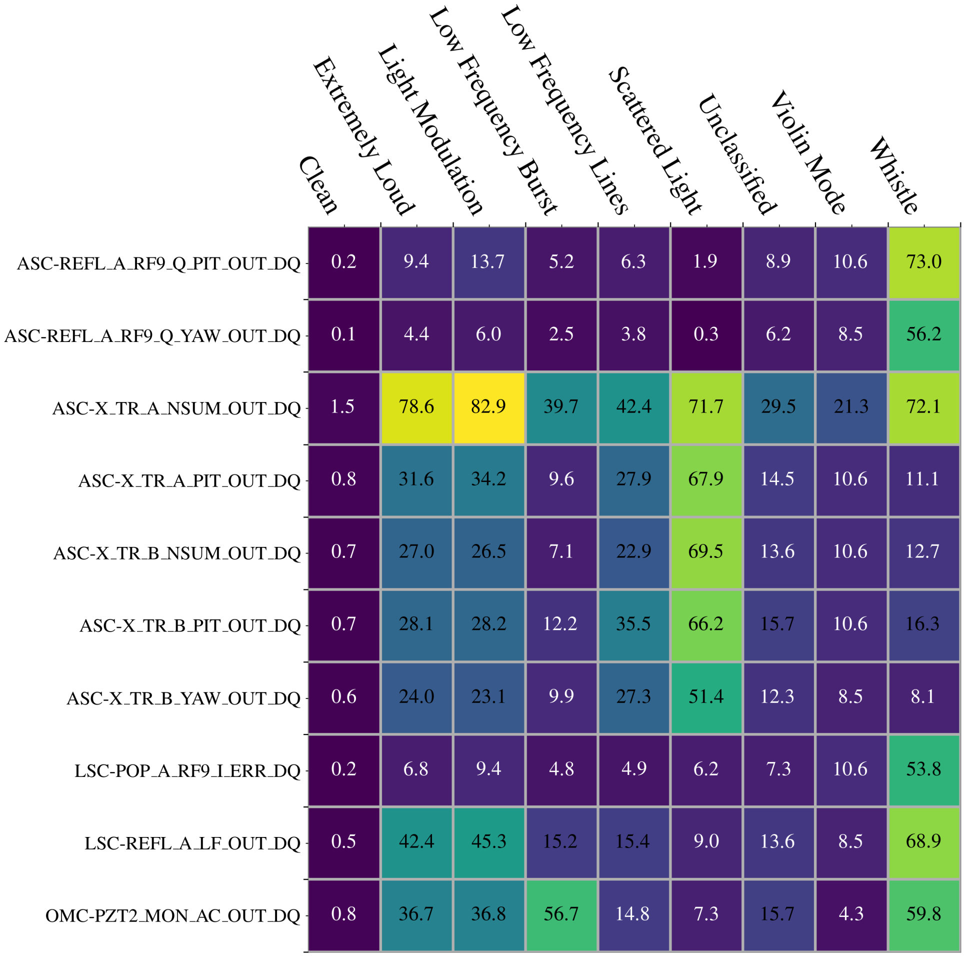
In this work, we've discussed improvements to the iDQ batch pipeline and we've demonstrated iDQ's ability to not only identify glitches in strain based solely on auxiliary channel behavior, but also shown its usefulness in identifying auxiliary channels which frequently report on the presence of glitches.

During O3b, it has been shown that iDQ had particularly powerful witness channels Scattered Light, Whistle, and Extremely Loud glitches. The correlation of these recovered events with the events that Omicron finds has been shown to be frequently two orders of magnitude greater than random chance, confirming that Omicron and iDQ are truly recovering the same events. Additionally, by analyzing the auxiliary channels alone, iDQ does not identify chirps, or likely real gravitational waves, to be glitches which Omicron reports in the same manner as any other glitch class. This not only proves the effectiveness of iDQ's identification scheme, but also the worth of its results alongside other glitch identification tools.

We have also introduced a method for calculating the probability that any time is a glitch based on iDQ data. We have seen that using the current method, the probability peaks at about 70%, or three orders of magnitude above probability given by Omicron identification alone. This demonstrates the power of combining results across multiple glitch identification tools, and could be a useful measure of data quality for inclusion in gravitational wave detection pipelines. The authors hope to implement this method in the near-future for use in real-time analyses.

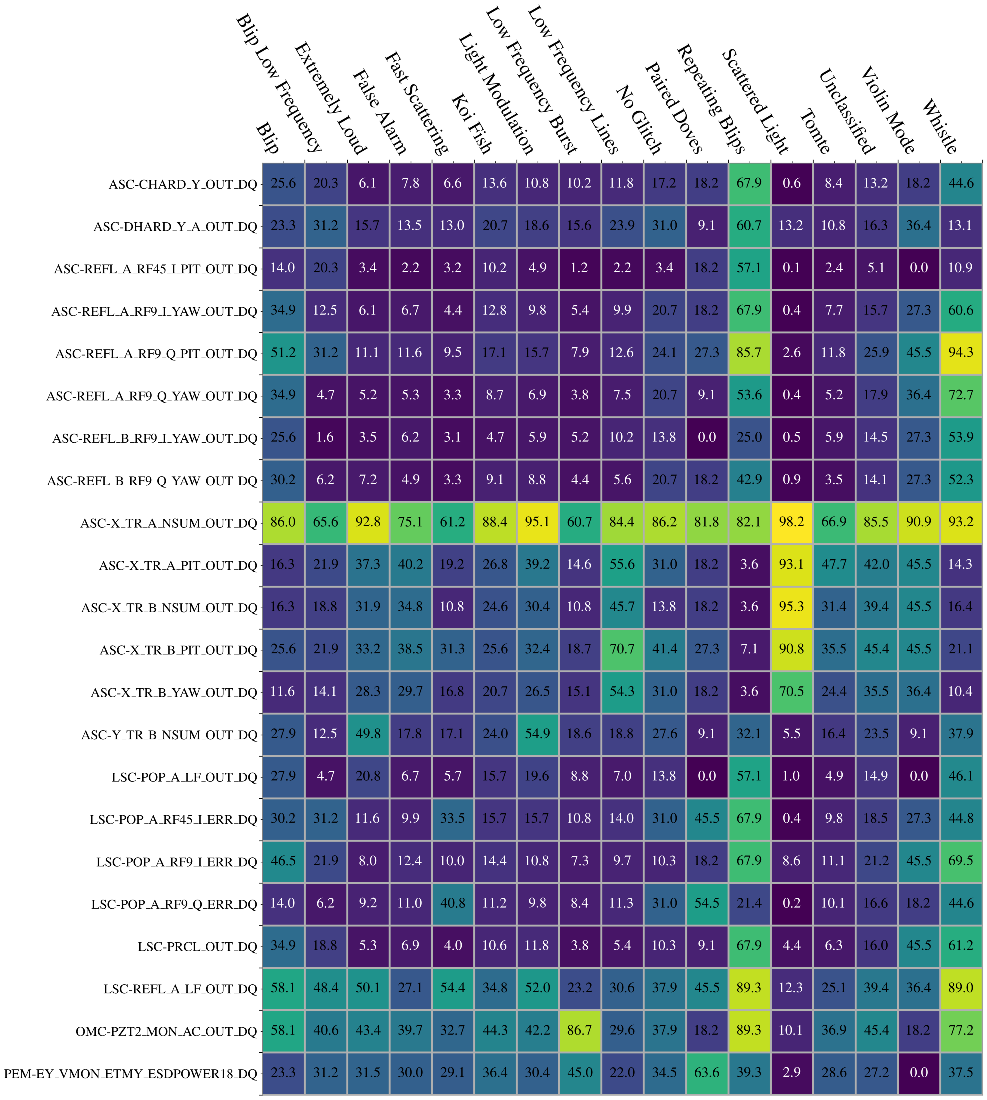
We have demonstrated the effectiveness in examining the correlations identified by iDQ between auxiliary channel activity and certain glitch classes. These correlations hint at possible sources of these glitch types, and when combined with follow-up from commissioners, could be a powerful tool in tracking down the origins of some glitch classes. Unfortunately, there is no way to know for sure whether any unique channel points to a true origin or whether a subsystem somewhere else in the detector is causing a glitch which then may propagate until witnessed by an unrelated auxiliary channel. Either way, channels like the ones we have mentioned could point to possible starting points for commissioners and detector characterization experts to begin looking for the sources of these extremely common glitches. If this analysis had been performed during an observing run, follow-up could have been done by commissioners on these channels, and these glitches could have been potentially mitigated during regular maintenance, or even during a longer commissioning break. The authors hope to make the auxiliary channel information available in low-latency in order to potentially impact maintenance and commissioning on the detectors moving forward.

As the detectors, glitch types, and auxiliary channels evolve between observing runs, iDQ will evolve with them, and the need for robust data quality information will only grow and sensitivity of the detectors increase. While this work demonstrates iDQ's past performance, it is heavily reliant on the quality of its auxiliary witnesses and as these change, so too will the types of glitches iDQ can identify and the efficiency at which it does so. Already, the LIGO Scientific and Virgo Collaborations have begun a fourth observing run in which the group has made many changes to the detectors and have a planned commissioning break before the second half in which more changes will be made. While results from the current observing run will of course vary the ones shown, iDQ has proven to be a reliable pipeline providing both probabilistic glitch identification, as well as glitch source identification, and it will continue to do so throughout O4 and beyond.



	Clean	Extremely Loud	Light Modulation	Low Frequency Burst	Low Frequency Lines	Scattered Light	Unclassified	Violin Mode	Whistle
ASC-REFL_A_RF9_Q_PIT_OUT_DQ	0.2	9.4	13.7	5.2	6.3	1.9	8.9	10.6	73.0
ASC-REFL_A_RF9_Q_YAW_OUT_DQ	0.1	4.4	6.0	2.5	3.8	0.3	6.2	8.5	56.2
ASC-X_TR_A_NSUM_OUT_DQ	1.5	78.6	82.9	39.7	42.4	71.7	29.5	21.3	72.1
ASC-X_TR_A_PIT_OUT_DQ	0.8	31.6	34.2	9.6	27.9	67.9	14.5	10.6	11.1
ASC-X_TR_B_NSUM_OUT_DQ	0.7	27.0	26.5	7.1	22.9	69.5	13.6	10.6	12.7
ASC-X_TR_B_PIT_OUT_DQ	0.7	28.1	28.2	12.2	35.5	66.2	15.7	10.6	16.3
ASC-X_TR_B_YAW_OUT_DQ	0.6	24.0	23.1	9.9	27.3	51.4	12.3	8.5	8.1
LSC-POP_A_RF9_I_ERR_DQ	0.2	6.8	9.4	4.8	4.9	6.2	7.3	10.6	53.8
LSC-REFL_A_LF_OUT_DQ	0.5	42.4	45.3	15.2	15.4	9.0	13.6	8.5	68.9
OMC-PZT2_MON_AC_OUT_DQ	0.8	36.7	36.8	56.7	14.8	7.3	15.7	4.3	59.8

Figure 7.8. Auxiliary channels that form vetoes in the associated OVL model at least 20% of the time at a log-likelihood of 2 or more for a variety of glitch classes. The included glitch classes are caught in coincidence by iDQ and Omicron at least ten times over the course of O3b with an iDQ log likelihood of at least 2 and an Omicron SNR of 10 or greater. Additionally included are the auxiliary channel results for the false alarms flagged by iDQ weighted by the total time not covered by Omicron glitches for comparison. Note that the majority of channels which are present in glitch veto lists, are not frequently present in clean veto lists.



	Blip	Blip Low Frequency	Extremely Loud	False Alarm	Fast Scattering	Koi Fish	Light Modulation	Low Frequency Burst	Low Frequency Lines	No Glitch	Paired Doves	Repeating Blips	Scattered Light	Tomte	Unclassified	Violin Mode	Whistle
ASC-CHARD_Y_OUT_DQ	25.6	20.3	6.1	7.8	6.6	13.6	10.8	10.2	11.8	17.2	18.2	67.9	0.6	8.4	13.2	18.2	44.6
ASC-DHARD_Y_A_OUT_DQ	23.3	31.2	15.7	13.5	13.0	20.7	18.6	15.6	23.9	31.0	9.1	60.7	13.2	10.8	16.3	36.4	13.1
ASC-REFL_A_RF45_I_PIT_OUT_DQ	14.0	20.3	3.4	2.2	3.2	10.2	4.9	1.2	2.2	3.4	18.2	57.1	0.1	2.4	5.1	0.0	10.9
ASC-REFL_A_RF9_I_YAW_OUT_DQ	34.9	12.5	6.1	6.7	4.4	12.8	9.8	5.4	9.9	20.7	18.2	67.9	0.4	7.7	15.7	27.3	60.6
ASC-REFL_A_RF9_Q_PIT_OUT_DQ	51.2	31.2	11.1	11.6	9.5	17.1	15.7	7.9	12.6	24.1	27.3	85.7	2.6	11.8	25.9	45.5	94.3
ASC-REFL_A_RF9_Q_YAW_OUT_DQ	34.9	4.7	5.2	5.3	3.3	8.7	6.9	3.8	7.5	20.7	9.1	53.6	0.4	5.2	17.9	36.4	72.7
ASC-REFL_B_RF9_I_YAW_OUT_DQ	25.6	1.6	3.5	6.2	3.1	4.7	5.9	5.2	10.2	13.8	0.0	25.0	0.5	5.9	14.5	27.3	53.9
ASC-REFL_B_RF9_Q_YAW_OUT_DQ	30.2	6.2	7.2	4.9	3.3	9.1	8.8	4.4	5.6	20.7	18.2	42.9	0.9	3.5	14.1	27.3	52.3
ASC-X_TR_A_NSUM_OUT_DQ	86.0	65.6	92.8	75.1	61.2	88.4	95.1	60.7	84.4	86.2	81.8	82.1	98.2	66.9	85.5	90.9	93.2
ASC-X_TR_A_PIT_OUT_DQ	16.3	21.9	37.3	40.2	19.2	26.8	39.2	14.6	55.6	31.0	18.2	3.6	93.1	47.7	42.0	45.5	14.3
ASC-X_TR_B_NSUM_OUT_DQ	16.3	18.8	31.9	34.8	10.8	24.6	30.4	10.8	45.7	13.8	18.2	3.6	95.3	31.4	39.4	45.5	16.4
ASC-X_TR_B_PIT_OUT_DQ	25.6	21.9	33.2	38.5	31.3	25.6	32.4	18.7	70.7	41.4	27.3	7.1	90.8	35.5	45.4	45.5	21.1
ASC-X_TR_B_YAW_OUT_DQ	11.6	14.1	28.3	29.7	16.8	20.7	26.5	15.1	54.3	31.0	18.2	3.6	70.5	24.4	35.5	36.4	10.4
ASC-Y_TR_B_NSUM_OUT_DQ	27.9	12.5	49.8	17.8	17.1	24.0	54.9	18.6	18.8	27.6	9.1	32.1	5.5	16.4	23.5	9.1	37.9
LSC-POP_A_LF_OUT_DQ	27.9	4.7	20.8	6.7	5.7	15.7	19.6	8.8	7.0	13.8	0.0	57.1	1.0	4.9	14.9	0.0	46.1
LSC-POP_A_RF45_I_ERR_DQ	30.2	31.2	11.6	9.9	33.5	15.7	15.7	10.8	14.0	31.0	45.5	67.9	0.4	9.8	18.5	27.3	44.8
LSC-POP_A_RF9_I_ERR_DQ	46.5	21.9	8.0	12.4	10.0	14.4	10.8	7.3	9.7	10.3	18.2	67.9	8.6	11.1	21.2	45.5	69.5
LSC-POP_A_RF9_Q_ERR_DQ	14.0	6.2	9.2	11.0	40.8	11.2	9.8	8.4	11.3	31.0	54.5	21.4	0.2	10.1	16.6	18.2	44.6
LSC-PRCL_OUT_DQ	34.9	18.8	5.3	6.9	4.0	10.6	11.8	3.8	5.4	10.3	9.1	67.9	4.4	6.3	16.0	45.5	61.2
LSC-REFL_A_LF_OUT_DQ	58.1	48.4	50.1	27.1	54.4	34.8	52.0	23.2	30.6	37.9	45.5	89.3	12.3	25.1	39.4	36.4	89.0
OMC-PZT2_MON_AC_OUT_DQ	58.1	40.6	43.4	39.7	32.7	44.3	42.2	86.7	29.6	37.9	18.2	89.3	10.1	36.9	45.4	18.2	77.2
PEM-EY_VMON_ETMY_ESDPOWER18_DQ	23.3	31.2	31.5	30.0	29.1	36.4	30.4	45.0	22.0	34.5	63.6	39.3	2.9	28.6	27.2	0.0	37.5

Figure 7.9. Auxiliary channels which appear in the top 10 vetoes of the associated OVL model at least 20% of the time for a variety of glitch classes which are caught in coincidence by iDQ and Omicron at least ten times over the course of O3b with an iDQ log likelihood of at least 2 and an Omicron SNR of 10 or greater. Additionally included are the auxiliary channel results from a random set of clean samples for comparison. Even at the higher log likelihood threshold cut-off, the majority of channels present witness some combination of glitch types, but mostly exclude clean samples. A sign that these channels are truly good witnesses of glitches, and not just frequently active.

Chapter 8
The Accuracy of Neutron Star Radius Measurement with the Next Generation of Terrestrial Gravitational-Wave Observatories

8.1 Introduction and Background

An outstanding problem in nuclear astrophysics is the equation-of-state of neutron star (NS) cores, believed to contain matter at several times the nuclear saturation density [179–181]: near the core the density reaches 4 to 6 times the nuclear saturation density and in the outer core it would be twice the nuclear saturation density. This makes them the densest objects anywhere in the Universe. Decades after their discovery, the radii of neutron stars are still uncertain[1] by about $\sim 10\%$ [182–188], and the composition of their dense cores likely depends on the neutron star mass and could be composed of hadrons or deconfined quarks [181, 189–192]. Indeed, it is not clear whether the matter at such densities undergoes a phase transition from a hadronic phase to quark-gluon plasma and the critical neutron star mass and temperature at which the transition might occur [181, 189, 190, 190–194].

Neutron stars in binaries are studied either as radio pulsars or X-ray sources and both have helped in our understanding of the structure of neutron stars [195–201]. The Neutron Star Interior Composition Explorer (NICER) space observatory is providing precision X-ray data on neutron stars [202]. Precise general relativistic modeling of the X-ray pulsation of neutron stars has been used to constrain their masses and radii as well as the equation-of-state (EOS) of their dense cores [182, 188, 203–207]. The best-measured NICER radius errors are about 1 km.

At the same time, advances in gravitational-wave observations from merging neutron stars are allowing new approaches to resolve this puzzle. Indeed, the detection of binary neutron stars (BNSs) [17, 208–211] and neutron star-black hole binaries (NSBHs) [212] by the Laser Interferometer Gravitational-Wave Observatory (LIGO) and Virgo has opened up a new and

[1] Note that some authors, who claim a 5% uncertainty in the radius, are quoting one-sided, one-σ credible intervals. The 10% to which we refer corresponds to a two-sided, 90% credible interval, which is the standard in LIGO-Virgo Collaboration publications.

independent window for exploring neutron stars. Gravitational waves emitted in the final tens of milliseconds of the inspiral and coalescence of BNSs can be used to explore the composition and EOS of dense matter in neutron star cores [213–217]. Encoded in the phase evolution of the waves is the (dimensionless) *tidal deformability* $\Lambda_{1,2}$ of the two stars, which is a measure of the quadrupole deformation imparted on the stars due to the tidal field of their companions. The leading order finite size effect in the post-Newtonian (PN) approximation of the waves' phase evolution is a highly sub-dominant effect. In terms of the post-Newtonian expansion parameter $(v/c) < 1$, it is, in fact, an order $O(v/c)^{10}$ effect beyond the dominant quadrupole term [216–219], yet it is significant when the instantaneous gravitational-wave frequencies are ~ 100 Hz or greater ($v/c \sim 0.16$ or larger) for a typical BNS system comprising a pair of 1.4 $M_\odot$ companions [220].

The tidal deformability goes as the inverse fifth power of the star's compactness, i.e. $\Lambda_k \propto [Gm_k/(c^2 R_k)]^{-5}$, $k = 1, 2$, where m_k and R_k are the masses and radii of the companion stars in a binary system [218,221]. Matched filtering the data with gravitational-wave templates calibrated to numerical relativity simulations [222–229] of BNS mergers can be used, in principle, to measure the tidal deformabilities of the companions, in addition to their masses[2]. In practice, however, it is not possible to accurately measure the individual tidal deformabilities, but only a certain linear combination of the two called *effective tidal deformability* $\tilde{\Lambda}$, defined by:

$$\tilde{\Lambda} = \frac{16}{13(1+q)^5}\left[(1+12q)\,\Lambda_1 + q^4\,(12+q)\,\Lambda_2\right] \tag{8.1}$$

where $q \equiv m_2/m_1 \leq 1$ is the mass ratio [216,219,221,230]. Although the dominant tidal effect, which depends only on $\tilde{\Lambda}$, is measured accurately, the PN correction, required to measure the individual tidal deformabilities, cannot be inferred with any accuracy. This is because of two reasons: On the one hand, it is a higher order PN correction, an $O(v/c)^{12}$ effect, compared to the dominant quadrupole term and, on the other hand, the PN correction vanishes for binaries with comparable masses. In fact, the tidal PN correction depends on $\delta\tilde{\Lambda}$ defined by:

$$\begin{aligned}\delta\tilde{\Lambda} &= \sqrt{1-4\eta}\left(1 - \frac{13272}{1319}\eta + \frac{8944}{1319}\eta^2\right)\left(\frac{\Lambda_2+\Lambda_1}{2}\right)\\ &+ \left(1 - \frac{15910}{1319}\eta + \frac{32850}{1319}\eta^2 + \frac{3380}{1319}\eta^3\right)\left(\frac{\Lambda_2-\Lambda_1}{2}\right),\end{aligned} \tag{8.2}$$

where $\eta \equiv m_1 m_2/(m_1+m_2)^2 = q/(1+q)^2$ is the symmetric mass ratio. For BNS systems in general, companion masses are similar, and hence $q \simeq 1$ and $\Lambda_1 \simeq \Lambda_2$, giving $\tilde{\Lambda} \simeq \Lambda_{1,2}$ and hence $\delta\tilde{\Lambda} \simeq 0$. Additionally, the tidal deformability of a neutron star depends not only on its mass, but also the (unknown) EOS. For neutron stars of $1.4 M_\odot$ and over a wide range of equations-of-state (EOSs), typical values are $\Lambda_k \sim 200$–2000 [221]. While the first post-Newtonian correction is already sub-dominant as a sixth post-Newtonian order effect compared to the leading order quadrupole [221], this range of Λ_k also results in the term being at least two orders of magnitude smaller compared to the leading order tidal term. These effects combined

[2] Neutron stars in merging binaries are not expected to have large spins. Consequently, the only intrinsic parameters that we will consider in this paper are the masses and the tidal deformabilities.

make the term difficult to measure. Consequently, only the leading order tidal term, is readily available, making it necessary to supplement gravitational-wave observations with other input in order to infer the individual tidal deformabilities and the radii of neutron stars. Several such approaches have been proposed in the literature and applied to GW170817 [210, 231].

The BNS coalescence event GW170817, at ~ 40 Mpc and a signal-to-noise ratio (SNR) of 33, provided the first opportunity to constrain the tidal deformabilities from gravitational-wave observations, and hence the radii, of neutron stars [17, 210, 211]. Theoretical models of the EOS of neutron stars are plenty and varied and they allow tidal deformabilities in the range of $10 \lesssim \Lambda_{1,2} \lesssim 10000$ [217, 218], depending on the mass, being larger for lighter neutron stars and stiffer EOSs. Analysis of the event GW170817 found that the 90% credible range of the companion masses were $1.36\,M_\odot \leq m_1 \leq 1.89\,M_\odot$ for the primary and $1.00\,M_\odot \leq m_2 \leq 1.36\,M_\odot$ for the secondary [17], the effective tidal deformability had a 90% credible upper bound of $\tilde{\Lambda} \lesssim 600$ and the radius was constrained to be $R_1 = 11.9^{+1.4}_{-1.4}$ km [210, 232]. Unfortunately, the second BNS event GW190425 [233] was farther and had a significantly lower SNR than GW170817 and did not yield tighter constraints on the tidal deformability on its own.

However, constraints have also been derived by combining LIGO-Virgo results of GW170817 and GW190425 with additional observations. Including NICER observations [182, 183, 188, 203–207] bound the radius of a $1.4\,M_\odot$ neutron star to the range $R_{1.4} = 12.33^{+0.76}_{-0.81}$ km. Likewise, combining nuclear physics experiments and gravitational-wave data has found $R_{1.4} = 11.0^{+0.9}_{-0.6}$ km [234], and $R_{1.4} = 12.75^{+0.42}_{-0.54}$ km [235] while combining data from GW170817, its companion gamma-ray burst GRB170817A, and subsequent kilonova AT2017gfo, the same radius was determined to an accuracy of less than about a km at 90% credible interval [236]. However, see [237] for sensitivity of NICER results on model hypotheses.

The planned upgrades of LIGO and Virgo, the addition of observatories currently under construction, KAGRA [36] in Japan and LIGO-Aundha in India [35], and new, longer-arm facilities that are currently being conceived, have the potential to make new discoveries of both sources and science. In this study, we explore the accuracy with which future observatories are able to measure the radii of neutron stars, an important step in constraining their equation of state. The networks considered in this work include the imminent upgrade of LIGO and Virgo over the next five years called A+ [208, 209], the *Voyager* upgrade to LIGO detectors that would be possible within the next ten years [34], and the next-generation (XG) observatories such as the Einstein Telescope (ET) [37–39] or Cosmic Explorer (CE) [40] that are expected to operate in the mid-2030s in tandem with the fully upgraded versions of current observatories. Given the rate of BNS mergers as determined by the events GW170817 and GW190425, we expect the future observations to constrain the neutron star radius to within 600 m (A+ generation), 400 m (Voyager generation), 200 m (one XG observatory) and < 100 m (two or more XG observatories), with the high-fidelity events observed by the respective networks of observatories. At the same time, neutron star masses will be measured to better than 10%, 5%, 3% and 0.5% [238]. The mass-radius relation is a proxy to the EOS of ultra-dense matter in neutron-star cores that will be tightly constrained with high-precision measurements of the masses and radii with future networks of gravitational-wave observatories (see, e.g., [239]).

When combining information from a multiple set of events it is necessary to employ a population model for the observed sources in addition to the unknown equation of state. For binary neutron stars, the population model will involve the astrophysical distribution of neutron star masses (or, equivalently, the neutron star central densities), the pairing probability as a function of the total mass and mass ratio and the distribution of neutron star spins. Moreover, gravitational-wave detectors and the analysis pipeline used to detect binary neutron stars have selection effects. For example, it is easier to detect equal-mass systems compared to mass-asymmetric systems of the same total mass. Likewise, binaries with a larger total mass produces a larger signal-to-noise ratio compared to a binary of smaller total mass but the same mass ratio. Bayesian inference of the source parameters for a single event will also be affected by the unknown hyper parameters of the population model since the posterior distribution depends on the assumed prior model. Thus, one has to simultaneously determine the population model and the EOS. For the EOS, this means one has to marginalize over the population model. Additionally one must also account for the selection effects to assure that the model selection of EOS is unbiased.

We are ignoring these effects in this work since our Fisher matrix approach currently does not allow for the inclusion of systematic biases. We also envisage that in the XG era the selection effects would have been better understood. Our goal, instead, is to provide the statistical uncertainty that we expect in the determination of the EOS. We are currently in the process of preparing a mock data challenge for XG observatories. The mock data challenge will allow us to address the aforementioned issues.

We also note that the estimation of intrinsic source masses requires the use of a cosmological model. Since we detect BNS events to a significant cosmological distance, cosmological parameters must be inferred together with the parameters of a BNS event [240, 241]. As explained in Sec. 8.5.5, we find that the bias introduced due to an unknown cosmological model is negligible.

The rest of the paper is organized as follows. In Sec. 8.2 we describe the cosmic BNS population used in this study together with the distribution of companion masses, the merger rate and its variation with redshift and the waveform model used. This is followed by a brief summary of detector networks considered in Sec. 8.3, focusing on the efficiency of the networks in detecting BNS systems. In Sec. 8.4 we present the capabilities of the different observatories in characterizing the source properties. We describe in Sec. 8.5 the method to infer the radii of neutron stars from the measurement of effective tidal deformability using a set of EOS independent universal relations with corrections and how we combine the results from a population to obtain joint bounds. In Sec. 8.6 we present the application of the methods to events expected to be observed in detector networks considered in this study. The results are obtained by combining radius measurements of a small sub-population of observed events: either the loudest 100 events or the 100 events for which tidal deformability is best measured, to infer the radii of neutron stars. A summary of the results and conclusions is presented in Sec. 8.7.

8.2 Neutron Star Population and Waveform Model

In this section, we describe the neutron star population and the waveform approximations used in the study. We begin by recalling how the redshift dependence of the merger rate is computed using the observed star formation rate as a function of redshift as a proxy for the redshift evolution of the rate. The redshift dependence is not exactly the same as the star formation rate since binaries that form from stars only merge after a certain time delay, which is essentially the gravitational radiation back reaction timescale. This is followed by a summary of the distribution of neutron star masses used in the study. We conclude the section with a description of the waveform model used, which is built upon the point-particle approximation but includes finite-size tidal effects with the waveform model parameters calibrated to hydrodynamical numerical relativity simulations of BNS mergers.

8.2.1 BNS Merger Rate

The merger rate density r_0 in the local Universe (i.e., at zero redshift) inferred from LIGO-Virgo observations of BNS coalescences during the second and third observing runs is $r_0 = 10 - 1700\ \mathrm{yr}^{-1}\ \mathrm{Gpc}^{-3}$ [6]. The two BNS events observed during this period, GW170817 and GW190425, were localized to luminosity distances of 40 Mpc and 159 Mpc, respectively, and corresponding redshifts of $z \simeq 0.01$ and $z \simeq 0.036$. Thus, the LIGO-Virgo rate is essentially the *local* merger rate density, i.e. at redshift $z = 0$. In this work, we will consider mergers up to a redshift of $z = 1$. The merger rate density over this redshift range is expected to increase since the rate of star formation $\psi(z)$, from which compact binaries form, increases with redshift up to about $z = 2$ [7].

To model the variation of the merger rate with redshift, we assume that it follows the star formation rate except that a binary that forms at redshift z_1 merges at redshift $z < z_1$. This is because there could be a significant time-delay t_d between the binary's initial formation and eventual merger as driven by gravitational radiation back reaction. The time delay t_d for a specific binary depends on a number of astrophysical processes that take place between the formation of the companion stars, their common evolution, and survival following supernova kicks they receive. Therefore, t_d will not be the same for every binary and the time delay distribution is not well known either due to the complexity of how the progenitors of compact binaries evolve. However, making reasonable assumptions about the intervening processes, i.e. neutron stars form with no delay after the formation of their progenitor stars, their orbit decays due to the emission of gravitational waves only, and the semi-major axis of their orbit follows a uniform in log-space distribution, t_d follows the distribution $P(t_d) \propto 1/t_d$ [242, 243]. Thus, the merger rate density in the source's frame[3] $r_z(z)$ is given by:

$$r_z(z) = A \int_{t_d^{\min}(z)}^{t_d^{\max}(z)} \psi(z - t_d(z))\, P(t_d(z))\, \frac{dt_d}{dz}\, dz, \tag{8.3}$$

where a subscript z is included to clarify that $r_z(z)$ is the rate density with respect to an observer at z, $t_d^{\min}$ and $t_d^{\max}$ are the minimum and maximum time delays, A is a normalization constant (see below), and $\psi(z)$ denotes the star formation rate (whose dimensions are not important to us but only its dependence on redshift). For $\psi(z)$ we use the fit proposed in Ref. [244]:

[3] In what follows lower case letters are used to denote the merger rate densities while capital letters are used to denote the merger rates.

$$\psi(z) \propto \frac{a \exp\left(b(z - z_m)\right)}{a - b + b \exp\left(a(z - z_m)\right)} \tag{8.4}$$

where $a = 2.8$, $b = 2.46$, and $z_m = 1.72$. For the minimum time-delay we use $t_d^{\min} = 0.2$ Gyr and for the maximum we use $t_d^{\max} = 10$ Gyr. The normalization constant A is determined so that this expression is consistent with the local rate density, i.e. $r_0(z = 0) = r_0$. The merger rate, $r_z(z)$, peaks at a slightly lower redshift than $\psi(z)$ because of the time-delay. The dependence of the cosmic time t on redshift is determined by the Planck 2015 Cold Dark Matter cosmology:

$$\frac{dt}{dz} = \frac{1}{H_0(1 + z)\,\sqrt{\Omega_\Lambda + \Omega_M(1 + z)^3}}, \tag{8.5}$$

with the Hubble constant $H_0 = 69.6\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, $\Omega_\Lambda = 0.714$ and $\Omega_M = 0.286$.

Next, the merger rate (as opposed to rate density) in a redshift interval dz is given by:

$$dR_z(z) = r_z(z)\,\frac{dV}{dz}\,dz \tag{8.6}$$

where $dV = (dV/dz)\,dz$ is the comoving volume element corresponding to redshift range dz. To convert this to the rate as measured by an observer at $z = 0$ we must divide by $(1 + z)$ to take into account the redshift of the rate due to cosmological expansion: $dR_0(z) = dR_z(z)/(1 + z)$. The cumulative merger rate is given by:

$$R(z) = \int_0^z dR(z') = \int_0^z \frac{r_z(z')}{(1 + z')}\,\frac{dV}{dz'}\,\mathrm{d}z'. \tag{8.7}$$

Within $z = 1$ the merger rate is about $\sim 10^5$ per year. Not all of these mergers would be detectable by a gravitational-wave detector (or a network) but only a certain fraction depending on its sensitivity which we will discuss in Sec. 8.3.

8.2.2 Waveform Models and Mass Distribution

In order to characterize the capability of various detector networks to measure the tidal deformability and the companion masses, it is important to choose an appropriate waveform model that includes the relevant physical effects. As in the case of binary black holes, BNS waveforms are based on approximate solutions to Einstein equations. They include the dominant tidal effects and incorporate additional parameters in the phase evolution which are calibrated by matching the analytical solution against numerical relativity simulations. We chose the frequency-domain phenomenological waveform model IMRPhenomPv2NRTidalv2 [224–226] for the generation of simulated signals as well as templates for Fisher-matrix based inference. This model is based on the IMRPhenomPv2 BBH waveform [245, 246] with tidal effects up to 7.5 post-Newtonian order (or to $O(v/c)^{13}$ beyond the leading quadrupole term), making it appropriate for use in BNS analysis. An earlier version of this waveform was used for the analysis of GW170817 [12].

The waveform model takes as input the intrinsic masses of the companions and their tidal deformabilities. In this paper, the companion masses are drawn from a uniform distribution over a range of masses whose lower limit is 1 $M_\odot$ and the upper limit is the maximum allowed by the EOS used in the simulation (see below): $m_1, m_2 \sim U(1\,M_\odot,\, M_{\rm max}^{\rm EOS})$. Although the masses of neutron stars in the Milky Way seem to be concentrated around 1.4 $M_\odot$, there is *a priori* no physical reason to assume that this is the preferred value in other galaxies. Theoretically, neutron star masses are allowed to be as large as 2.9 $M_\odot$ [247], although the largest measured masses tend to be significantly lower. The heaviest neutron stars among astronomical observations are in range 2.01–2.35 $M_\odot$ [248–250], while from gravitational-wave observations the companion masses in BNS systems are as large as 1.6 $M_\odot$, and 1.4 $M_\odot$ in the case of GW170817, and 1.9 $M_\odot$, and 1.7 $M_\odot$ in the case of GW190425. Neutron-star masses in neutron star-black hole systems GW200105 and GW20015 [59] are both 2.2 $M_\odot$. In this small population there seems to be no preference for the Galactic value of $\sim 1.4\,M_\odot$ and it would be more prudent to assume a wider range for the mass distribution. We have chosen the widest range allowed by the model EOSs considered in this paper.

We assume, however, that the dimensionless spin magnitudes of neutron stars are negligible. The fastest-spinning Galactic pulsar has a rotational frequency of just over 700 Hz. Its dimensionless spin angular momentum is still roughly $a = cI2\pi\omega/Gm^2 \simeq 0.4$—far smaller than the maximum spin neutron stars could, in principle, have; here I is the principal moment-of-inertia of the star (roughly equal to $\frac{2}{5}mR^2$, where m and R are the neutron star's mass and radius, respectively), and ω is its spin angular frequency. Neutron star spins in other galaxies could be far greater than those in the Milky Way but the waveform models that are currently available are calibrated against numerical relativity simulations of BNSs with small spins (dimensionless spin, $\chi < 0.1$) [224–226].

In addition to masses and spins, we also have to specify the distance to the source, its orientation relative to the detector frame, and its position in the sky. Sources are assumed to follow the redshift distribution determined by Eq. (8.6) and uniformly distributed over the angular parameters describing the sky position and orientation of the binary.

Given the mass, the radius of the neutron star is calculated for a given EOS by solving the Tollmann-Oppenheimer-Volkoff (TOV) equations [251,252]. In practice, this is computationally too expensive since our simulations have to deal with hundreds of thousands of systems. Thus, it is more practical to solve the TOV equations to obtain radii for a set of masses and then use an interpolating function to find the radius for an arbitrary value of mass. We have confirmed that the fractional difference in the radius, for a given mass, obtained from numerical solution to the TOV equation and the interpolating function are below 0.1% over the full range of neutron star masses allowed by the EOS.

We consider three EOS used for injection, and an additional seven EOS used for reference that are still allowed by X-ray and gravitational-wave data: the injection set of ALF2, APR3, APR4 and reference set of DD2, H4, S220, PP2, PP5, SFHo, and SLy. We then plot the corresponding mass-radius curves in Fig. 8.1. ALF2 (APR4) represents a stiffer (softer) EOS allowing for larger (smaller) radii, while APR3 allows intermediate radii. The reference EOS then provide good coverage of the mass-radius parameter space between the three, allowing for stronger model discrimination tests with our methods. Given the mass M_i and radius R_i, the

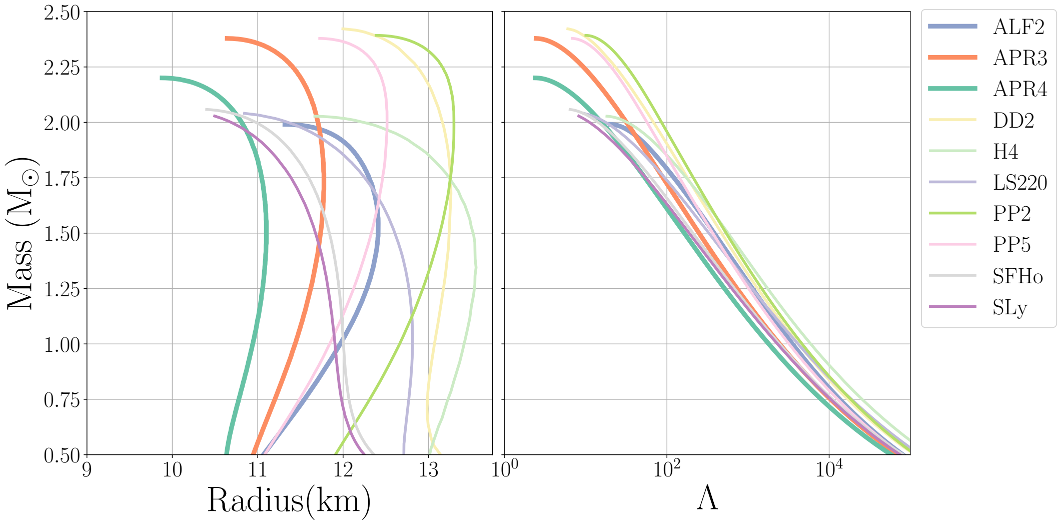


Figure 8.1. Mass-Radius curves for EOS used in this paper. Please note that our choice of three injections EOS here (ALF2, APR3, APR4 shown with thicker lines) are motivated by the conservative constraint on $\Lambda_{1.4} < 800$ as put forward by [12]. We also consider the fact that these three EOS covers a significant range in the maximum masses while the inclusion of addtional seven EOS provide good coverage of the rest of the mass radius space.

dimensionless tidal deformability is computed using the expression:

$$\Lambda_i = \frac{2\,k_2(R_i)}{3}\left(\frac{c^2 R_i}{G m_i}\right)^5, \tag{8.8}$$

where $k_2(R)$ is the tidal Love number, which also depends on the radius of the neutron star and is fixed for a given mass and EOS [218].

8.3 Future Observatories and Their Reach for the Binary Neutron Star Population

Advanced LIGO (aLIGO) and Advanced Virgo (AdV) are currently taking data and are expected to reach their design sensitivity goals (see Fig. 8.2) in late 2023[4] [208]. At that sensitivity, the network of LIGO-Hanford, LIGO-Livingston, and Virgo (HLV) [209] could detect ~ 40 BNS mergers per year from within a distance of about 400 Mpc. Both projects have concrete plans to upgrade their sensitivity over a period of two years, which we will refer to as the HLV+ network, enhancing the detection rate by about a factor ~ 5 by about 2027[4].

[4] For up to date schedule of the runs see `https://rtd.igwn.org/projects/userguide/en/latest/capabilities.html`.

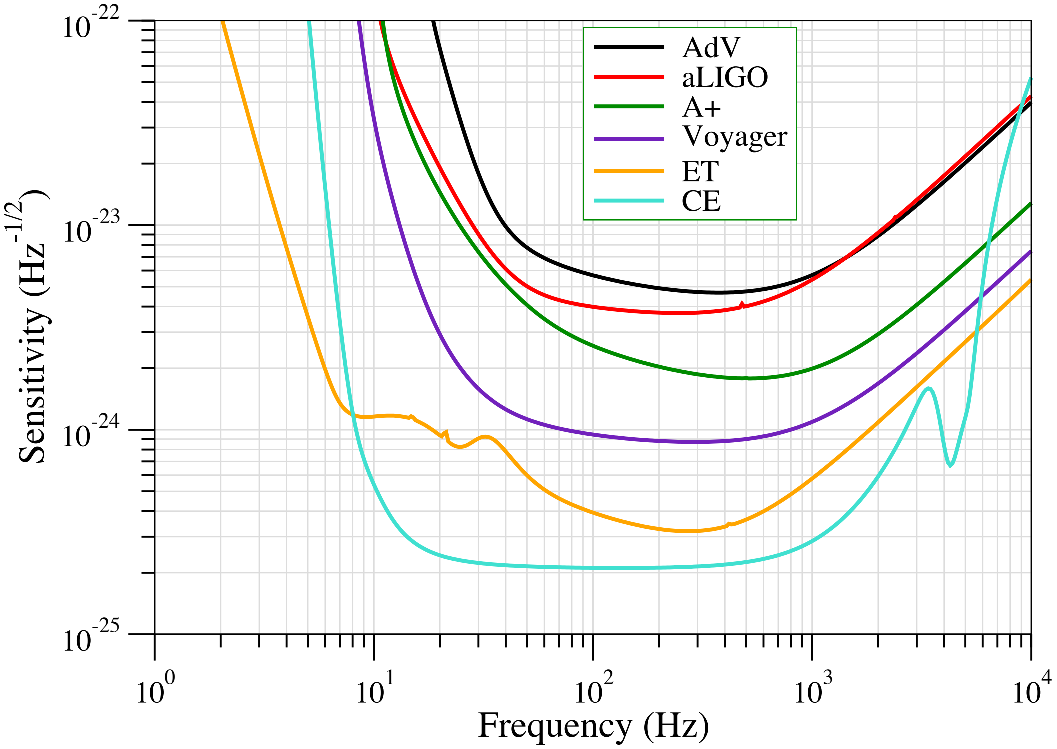


Figure 8.2. Strain sensitivity of three generations of ground-based gravitational wave detectors: (i) Advanced Virgo (AdV), Advanced LIGO (aLIGO) and A+, (ii) Voyager, and (iii) Einstein Telescope (ET) and Cosmic Explorer (CE). In the case of ET the sensitivity shown is that of an L-shaped detector with 10 km arms. The three V-shaped arms make the effective strain sensitivity a factor 3/2 better (and the noise floor lower by the same factor).

8.3.1 Upgrades and New Facilities

The Japanese KAGRA detector, currently being commissioned, and LIGO-India are expected to join the HLV+ network over the 2020-2030 decade and the five detectors would be together referred to as the HLVKI+ network. HLV+ and HLVKI+ networks will begin to observe events with SNRs large enough to facilitate accurate measurement of the tidal deformability.

Further upgrades to LIGO beyond A+ have been studied and they involve the development of new technology to mitigate thermal noise and gravity gradient background. Voyager [253] is one such concept that could lead to a further increase in the sensitivity by a factor of ~ 2–5 over the frequency range 10 Hz to a few kHz (see Fig. 8.2. At the moment we are not aware of any plans to upgrade Virgo or KAGRA and hence we will consider a network of five detectors: the three LIGO detectors operating with Voyager technology and Virgo and KAGRA in A+ mode. We will refer to this as the Voyager network, which will have access to several loud binary merger events. The Voyager network could constrain neutron star radius to within about 5% or roughly 500 m for neutron stars between $1.5M_\odot$ and $2.0M_\odot$, as seen in Figure 8.8.

Improvements in sensitivity beyond the level of Voyager would require, among other technologies, longer arms and/or underground facilities, neither of which would be possible

Table 8.1. Upgraded and future gravitational-wave detectors whose ability to measure the EOS of matter in neutron star cores is evaluated in this study. The time-scale of operation of the various networks is our best guess estimate of when a given network is likely to operate; they do not correspond to any official projections.

Detectors	Network Name
LIGO (HLI+), Virgo+, KAGRA+	HLVKI+
LIGO (HLI-Voy), Virgo+, KAGRA+	VK+HLIv
ET, LIGO (HLI+), KAGRA+	HLKI+E
CE, Virgo+, KAGRA+ LIGO-I+	VKI+C
ET, CE, KAGRA+, LIGO-I+	KI+EC
ET, CE, CE-South	ECS

with the infrastructure that exists at the location of current detectors. The boldest of the new concepts are the Einstein Telescope (ET) in Europe and Cosmic Explorer (CE) in the US and, possibly, Australia. ET is an underground facility hosting three V-shaped detectors at the vertices of an equilateral triangle of 10 km sides [39], while CE is a over-ground, L-shaped detector with 40 km arms [40]. ET and CE will be roughly 10 to 30 times more sensitive than advanced detectors (cf. Fig. 8.2) with the capability to observe hundreds of thousands BNSs mergers each year, many with SNRs larger than 100.

8.3.2 Detector Networks

Advanced LIGO, Advanced Virgo and KAGRA (LVK) have been taking data, albeit intermittently, since 2015, 2017 and 2019, respectively. They are expected to operate at design sensitivity during 2023-2024. We have not included the measurement capability of this network as the number of loud (i.e., SNRs in excess of 25) BNS coalescences expected to be detected during the next science run (O4) is only $\sim$ few.

LIGO-India, currently under construction, could join the upgraded A+ versions of the LVK network in the latter half of this decade; we shall call this the HLVKI+ network. Both LIGO and Virgo are planning for a further upgrade beyond 2030, referred to as Voyager in the US. A network in which Virgo and KAGRA operate at A+ sensitivities and LIGO-Hanford, LIGO-Livingston and LIGO-India operate at Voyager sensitivity, will be called VK+HLIv. This network will have the same performance as the one in which any three of 5 detectors are upgraded to Voyager and the remaining two operate at A+ sensitivity and we do not consider them separately.

Beyond 2035 one or more next generation observatories could begin to operate. To understand the relative merits of operating one or more such observatories we consider four different networks in which a subset of the current detectors operate at A+ sensitivity at the same time as one CE (which we shall denote VKI+C), one ET (denoted HLKI+E), one each of CE and ET (denoted KI+EC) and a network consisting of one ET, one CE in the US and one CE in Australia (denoted ECS) without any A+ detectors. In all, we consider six networks as enumerated in Table 8.1. For the ET and CE, we use fiducial locations and orientations as given in Ref. [254]. We will next discuss the expected performance of various detector networks in detecting signals from and measuring the parameters of BNSs.

8.3.3 Network Efficiency

Gravitational wave detectors have a wide field of view of the sky but they are not equally sensitive to all directions. An interferometric detector like LIGO has a quadrupole antenna pattern and is able to detect only a fraction of all the sources from within a given distance. A network of non-collocated detectors increases the sky coverage and the five-detector network of HLVKI+ has an almost isotropic response.

The efficiency of a detector network is a function of the luminosity distance (or redshift) and is defined as the fraction of all sources within a certain luminosity distance that can be (confidently) detected by the network, say with an SNR above a threshold SNR. In order to compute the efficiency of a detector network we simulate BNS events with their parameters distributed as described in Sec. 8.2.2. The network SNR of an event is simply the quadrature sum of the SNRs in each detector:

$$\rho^2 = \sum_{A=1}^{n_D} \rho_A^2, \quad \rho_A^2 = 4 \int \frac{|\tilde{h}^A(f)|^2}{S_h^A(f)}\, df, \tag{8.9}$$

where $\tilde{h}^A(f)$ is the Fourier transform of the response of detector A to an incident gravitational wave [cf. Eq. (8.12)], $S_h^{(A)}(f)$ is the one-sided noise power spectral density of detector A as in Fig. 8.2, ρ_A is the matched filter SNR of the signal in detector A, n_D is the number of detectors in the network, and ρ is the network SNR. The efficiency of a detector is then defined as:

$$\epsilon(z) = \frac{1}{N} \sum_k \Pi(\rho_k(z) - \rho_T), \tag{8.10}$$

where N is the total number of simulated events, $\rho_k(z)$ is the network SNR for the k$^{\text{th}}$ event, ρ_T is the SNR threshold and Π is the step function, $\Pi(x) = 0$, if $x < 0$ and $\Pi(x) = 1$, if $x > 0$. The SNR of an event depends not just on the redshift but on all other parameters of the source. In computing the network efficiency, we bin the SNR by redshift and ignore its dependence on all other parameters. The SNR threshold ρ_T serves as a proxy for detection confidence, larger SNRs are generally detected with greater confidence. We choose the threshold to be $\rho_T = 12$—the minimum SNR required for a network of detectors to make a confident detection. While the SNR of 12 used here is required for a confident detection, it is not necessarily the SNR at which we can make the accurate measurements of tidal deformability necessary to determine a neutron star's radius and its EOS. In later sections, we will choose the best subset of all events to evaluate how well a network is able to measure the radii of neutron stars.

The efficiency of a network then also determines its detection rate. Within a given redshift, a network does not observe all the possible sources, but only a fraction D_R given by:

$$D_R = \int_0^z \frac{r_z(z')}{(1+z')} \frac{dV}{dz'}\, \epsilon(z')\, dz'. \tag{8.11}$$

We call D_R the *detection rate* of a network and it is essentially the same as Eq. (8.7) except that the integrand is weighted with the efficiency of the network.

Table 8.2 lists the number of events detected over a period of **two** years, as a function of detection threshold. An SNR of 12 is required for a confident detection, and at that level, the A+ network would observe about 800 sources over two years while the Voyager network would observe almost ten times as many. Meanwhile, a network containing at least one XG detector would observe about half all the sources within $z = 1$, (70,000 if XG is ET and 100,000 if XG is CE) (see Table 8.2), and a network containing one ET and one CE would observe 30% more sources than that. The ECS network would additionally observe about 10% more sources than a network containing two XG detectors and 50% more than a network containing a single XG detector.

8.4 BNS Measurement Capability of Future Detector Networks

In this Section we assess the measurement capabilities of different networks of gravitational-wave detectors introduced in Sec. 8.3. We begin with a brief discussion of the distribution of the SNR in various detector networks followed by the accuracy with which parameters can be measured, in particular the effective tidal deformability.

In the rest of the paper, we will only consider sources up to a redshift of $z = 1$. Within this redshift, we expect about 150,000 BNS mergers over a two-year period but the current rate uncertainty means this number could be 50% larger or 25% smaller. This is a redshift that is far greater than the horizon distance of A+ and Voyager networks while a network containing one or more of XG detectors would observe a vast majority of mergers within it. However, only a small fraction of them will have large enough SNRs to be useful for measuring the EOS.

8.4.1 Signal to Noise Ratio Distribution for Nearby BNS Mergers

Figure 8.3 plots the cumulative distribution of the SNR for the population of BNS mergers up to a redshift of $z = 1$. The VKI+C network should observe 10% of the events with SNRs greater than 30 and 1% of the events with SNRs greater than 60. In contrast, in the A+ network less than 0.1% of events will have SNRs greater than 10. Cosmic Explorer and its southern counterpart operating along with Einstein Telescope would observe thousands of events each two years with SNRs greater than 100.

One must multiply the expected number of mergers within this redshift with the corresponding value of the CDF to get the number of sources expected to be observed each year. An estimate of actual number of events along with their SNR distribution is also given in Table 8.2.

8.4.2 Fisher Information Approach for Measurement Accuracy

Our goal is to estimate the accuracy with which parameters of an event can be measured by gravitational-wave detector networks. To this end, we employ the Fisher information matrix

ρ_T	HLVKI+	VK+HLIv	HLKI+E	VKI+C	KI+EC	ECS
12	840	7400	67,000	100,000	130,000	146,000
30	50	600	10,000	25,000	40,000	65,000
50	10	100	2,500	8000	12,000	23,000
100	0	10	300	1000	1,800	3800
300	0	0	10	50	70	150
500	0	0	1	5	10	30

Table 8.2. We list the number of events expected to be detected as we increase the SNR of events. Even with one Cosmic Explorer and/or Einstein Telescope, the number of BNS detections increases by an order of magnitude. In the bulk of this work, we focus our analysis on top 100 events with the highest SNR for each detector network. This cut corresponds to SNR of 100 or more for networks with at least on XG-era detector and about 50 or below for A+ detectors.

approach [255], which allows a reliable estimation of errors when the SNRs large (say more than about 30 or 50). We use the open source software GWBENCH [254] to generate and sample posteriors for a set of randomly selected signals. GWBENCH is a software package that computes the Fisher information matrix (FIM) $\mathcal{F}$ whose inverse gives the variance-covariance matrix. The starting point of the computation is the response of a detector to incident gravitational wave with polarizations h_+ and $h_\times$:

$$h^A(t, \boldsymbol{\theta}) = F_+^A(t, \alpha, \delta, \psi) h_+(t, \boldsymbol{\mu}) + F_\times(t, \alpha, \delta, \psi) h_\times(t, \boldsymbol{\mu}) \tag{8.12}$$

where A is an index denoting the detector in question. Here $F_{+,\times}$ are the plus and cross antenna pattern functions of the detector that depend on the right ascension α and declination δ of the source, and the polarization angle ψ. The time dependence of the antenna pattern functions are only important when the motion of the detector relative to the source is perceptible, and for sources that last for more than 30 minutes. The polarization amplitudes h_+ and $h_\times$ depend on the intrinsic parameters of the sources such as the masses m_1 and m_2 of the companion stars, [5] and the effective tidal deformability $\tilde{\Lambda}$, but also the extrinsic parameters that include the orientation ι of the binary's orbit relative to the line-of-sight from the Earth to the source and the source's luminosity distance D_L. These are all combined in the parameter $\boldsymbol{\mu} = \{\mathcal{M}, \eta, \tilde{\Lambda}, \iota, D_L\}$, where instead of the companion masses we have used the symmetric mass ratio $\eta \equiv m_1 m_2 / M^2$, and the chirp mass $\mathcal{M} \equiv \nu^{3/5} M$ ($M \equiv m_1 + m_2$). The parameter set $\boldsymbol{\theta}$ captures all the parameters describing the response of a detector to an incoming gravitational wave (see below for the full list of parameters).

Given the Fourier domain representation $h^A(f; \vec{\theta})$ of the detector response, the Fisher matrix is given by:

$$\mathcal{F}_{ij}^A = \left\langle \frac{\partial h^A(f)}{\partial \theta^i}, \frac{\partial h^A(f)}{\partial \theta^j} \right\rangle, \tag{8.13}$$

[5] In principle the companions can have spin angular momenta, but neutron stars are not expected to have large spins and they are not included in this study.

where the inner product of any two functions $a(f)$ and $b(f)$ is defined as

$$\langle a(f), b(f)\rangle = 2\int_{f_{\rm low}}^{f_{\rm high}} \frac{a(f)^* b(f) + a(f) b(f)^*}{S_h^A(f)} df. \tag{8.14}$$

where $a^*(f)$ denotes the complex conjugate of $a(f)$. The Fisher matrix of a network of detectors is simply the sum of the matrices corresponding to individual observatories in the network, i.e.

$$\mathcal{F}_{ij} = \sum_A \mathcal{F}_{ij}. \tag{8.15}$$

Given the Fisher matrix, the covariance matrix C_{ij} among the parameters is the inverse of the Fisher matrix, i.e. $C_{ij} = \mathcal{F}_{ij}^{-1}$.

To construct the Fisher likelihood surface, we choose a low-frequency cutoff, $f_{\rm low}$, of 10 Hz for A+ and Voyager detectors and 5 Hz for XG detectors. The high-frequency limit is taken to be the maximum allowed frequency given the sampling rate (typically chosen to be 4096 Hz), but the signal model never extends to such high frequencies even for the lowest-mass neutron stars considered in this paper. We then compute a 10-dimensional Fisher likelihood consisting of the parameter set $\boldsymbol{\theta} = \{\mathcal{M}, \eta, \tilde{\Lambda}, D_L, \psi, \cos\iota, \alpha, \delta, \phi_c, t_c\}$, where t_c, and ϕ_c are the fiducial time of coalescence, and the gravitational-wave phase at coalescence, respectively.

8.4.3 Measurement Accuracy of Simulated Population

Fig. 8.3 plots the errors on the parameters of the simulated population in the form of distribution functions. We have shown the results for a subset of all the parameters that are relevant to the measurement of the mass-radius curves. These are the chirp mass $\mathcal{M}$, the symmetric mass ratio η and the effective tidal deformability $\tilde{\Lambda}$. We see a clear delineation in the measurement capabilities of current and upgraded networks and XG observatories. The precise measurement of the parameters is, of course, accomplished by tracking the phase evolution of the binary. The chirp mass and mass ratio are most accurately measured if the number of cycles in the band is large (i.e. if the signal's phase can be tracked over longer periods) and a good improvement in low-frequency sensitivity for XG detectors is responsible for this vast improvement in the measurement of the mass parameters. The reduced tidal deformability measurement comes from the signal's phase evolution close to merger, or the high-frequency part of the signal, which will be clearly visible in XG detectors.

The remaining parameters—sky position, distance, and orientation of the binary in the plane of the sky—also show a clear delineation between detector generations, except the instance where the addition of CE without ET performs similar to Voyager networks[6].

8.4.3.0.1 Sky localization For very short transient signals, the sky localization is measured using the gravitational-wave travel times between different detectors and, therefore, depends on the number of non-collocated detectors. Thus, the 5-detector network of VK+HLIv, achieves

[6] The performance equivalence argued here is for a fraction relative to the total number of detected events. In absolute terms, even a single CE will have outstandingly more events with a given measurement error.

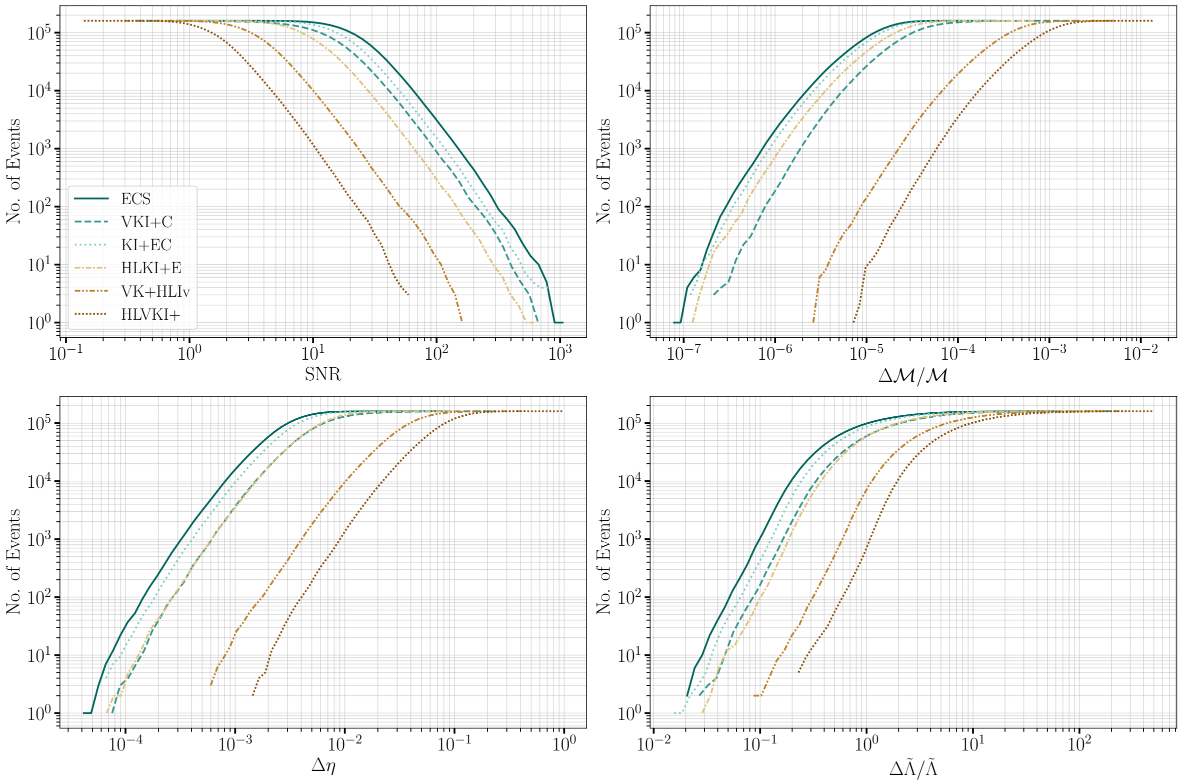


Figure 8.3. This plot shows the distribution of the measurement accuracy of the chirp mass $\mathcal{M}$, combined tidal deformability $\tilde{\Lambda}$, symmetric mass ratio η, and the SNR for 160 000, events expected over a two year period, up to a redshift of $z = 1$. The source parameters are distributed as described in Sec. 8.2.2.

greater precision than a 4-detector XG network VKI+C, although the signal strengths in the latter are much greater. For longer signals that make a discernible trail on the sky, the variation of the antenna response across the sky can be used to improve the sky position of the source. Since ET is more sensitive between 5 Hz and 8 Hz, where a typical BNS signal (1.4 $M_\odot$ + 1.4 $M_\odot$) spends more than an hour (~ 75 minutes), a trail spanning more than 15° on the sky (or, a fifth of the total variation in the antenna pattern) is clearer in the presence of an ET detector. Moreover, HLKI+E is composed of five detectors, which accentuates the sky resolution.

8.4.3.0.2 Inclination angle The measurement of the inclination angle is dependent on the distinguishability of the two gravitational-wave polarizations. Since ET is a triangular detector that measures three independent strains, each strain has different polarization content, leading to an accurate estimate of the polarization content and, thereby, the inclination angle. A CE detector alone cannot distinguish between the two gravitational-wave polarizations and it is the 2G background (inclined with respect to each other and CE) that provides crucial assistance to the VKI+C network in the polarization measurement. However, a mutually inclined 5-detector network VK+HLIv still achieves greater precision than a 4-detector VKI+C network.

8.4.3.0.3 Luminosity distance The luminosity distance parameter is most correlated with the inclination angle. Hence, a precise measurement of the inclination angle also leads to an accurate measurement of the luminosity distance. Thus, the measurement trends for the luminosity distance across networks follows the trends in the inclination angle.

8.5 Inferring Neutron Star EOS from Mass-Radius Curves

The Bayesian inference of the chirp mass $\mathcal{M}$, and symmetric mass ratio η of the BNS events detected by LIGO and Virgo are the most precise measurements among all parameters of BNS events. While the effective tidal deformability is not measured as precisely, upcoming gravitational-wave detector networks promise vastly improved measurements (cf. Fig.8.3). To measure the radii of component stars, however, it is necessary to know what the individual tidal deformabilities Λ_1 and Λ_2 are as well as the tidal Lover number k_2 (cf. Eq. 8.1). Unfortunately, gravitational-wave observations can only provide a reliable estimation of the linear combination $\tilde{\Lambda}$. This problem has been resolved temporarily via the proposal of a set of quasi-universal relations for neutron stars, which are approximately obeyed by hundreds of current models of the EOS [256].

In this set, there are basically two universal relations. The first of these relates the asymmetric combination of the individual tidal deformabilities[7] $\Lambda_a \equiv (\Lambda_2 - \Lambda_1)/2$ to the symmetric combination $\Lambda_s \equiv (\Lambda_2 + \Lambda_1)/2$ via the mass ratio q

$$\Lambda_a = F_n(q)\Lambda_s \frac{a + \sum_{i=1}^{3}\sum_{j=1}^{2} b_{ij} q^j {\Lambda_s}^{-i/5}}{1 + \sum_{i=1}^{3}\sum_{j=1}^{2} c_{ij} q^j {\Lambda_s}^{-i/5}}, \tag{8.16}$$

where the function $F_n(q)$ is given by

$$F_n(q) = \frac{1 - q^{10/(3-n)}}{1 + q^{10/(3-n)}}. \tag{8.17}$$

The fitting parameters b_{ij}, c_{ij}, a and n are given in Table I of Ref. [257]. The second universal relation [258] relates the compactness $C \equiv GM/(c^2 R)$ of an individual neutron star to its tidal deformability:

$$C(\Lambda) = \sum_{k=0}^{2} a_k (\ln \Lambda)^k, \tag{8.18}$$

where the fitting parameter a_k are also given in Table I of Ref. [257] (also see Ref. [259] for similar relationships).

The first of the universal relations Eq. (8.16) can be used to decouple the effective tidal deformability into individual tidal deformabilities. Then the second universal relation Eq. (8.18) can be used to compute the radius. These universal relations, however, have been shown to introduce systematic errors [260] that must be corrected in order to obtain an unbiased estimation of the EOS [261]. In the rest of this section, we describe our simulation method to assess the radii measurements for a set of future gravitational-wave observatories with corrections for these errors.

[7] We follow the convention $m_1 > m_2$ and, consequently, $\Lambda_1 < \Lambda_2$.

8.5.1 From Gravitational Wave Measurements to Neutron Star Radii

We begin with the Fisher information matrices (FIM), computed using the GWBENCH software, for the entire simulated BNS population and all the detector networks described in Sec. 8.2 for a set of three EOS models and the IMRPHENOMPV2NRTIDAL waveform model. Diagonal elements of the covariance matrix (inverse of the FIM) are the standard deviations of the source parameters: $(\mathcal{M}, \eta, \tilde{\Lambda}, \phi_c, t_c, D_L, \cos\iota, \alpha, \delta, \psi)$. In order to obtain radii of the companion stars from the parameters measured via gravitational-wave observation, we simulate posterior samples by generating a multi-dimensional Gaussian sample using the injection values as mean values and the inverse of the FIM as the covariance matrix. We need only three of these parameters $(\mathcal{M}, \eta, \tilde{\Lambda})$ for the estimation of radii. To break the degeneracy between two tidal deformabilities and get individual radii, we follow the procedure described in [261] (see also [12] for an alternative method), which is briefly described below.

First, in the expression for $\tilde{\Lambda}$ we eliminate Λ_1 and Λ_2 in terms Λ_s and Λ_a. We then use the first universal relation in Eq. (8.16) to replace Λ_a with Λ_s in the expression for $\tilde{\Lambda}$, thereby writing $\tilde{\Lambda}$ as a function of only Λ_s and q. Since gravitational-wave observations measure $\tilde{\Lambda}$, we can invert the expression for $\tilde{\Lambda} = \tilde{\Lambda}(\Lambda_s, q)$ to get $\Lambda_s(\tilde{\Lambda}, q)$. Thus, from gravitational-wave measurements of the mass ratio and the effective tidal deformability we can extract the symmetric combination Λ_s and then, using Eq. (8.16), also Λ_a. These two are then inverted to obtain the individual tidal deformabilities of the component stars. Thereafter, we use the C-Λ universal relation in Eq. (8.18) to derive the compactness and, with the individual masses, obtain the posterior probability distribution of the radii for component neutron stars.

8.5.2 Correcting Systematic Errors in Neutron Star Radii

Universal relations introduce systematic errors in the estimation of individual tidal deformabilities and radii which will dominate the source of errors in the era of XG observatories [261]. Due to the fact that $\delta\tilde{\Lambda}$ cannot be measured accurately, it is not possible to obtain a truly, arbitrarily precise, model-agnostic measurement of neutron star radii or compactness using only gravitational-wave measurements[8]. However, it turns out that for the purpose of EOS model selection the systematic errors can be corrected as we will briefly argue below (see Ref. [261] for details).

As discussed before, the GWBENCH framework is used to create a population of BNS events in which the tidal deformability Λ of each neutron star of mass m is computed for a specific EOS model (one of ALF2, APR3 or APR4). Out of the 150,000 simulated events, we choose 100 events that have either the greatest SNR or the best-measured tidal deformability. For the 100 events, we will have 200 mass-radius posteriors, one for each of the companion stars. We then sample a discretized mass-radius curve containing 200 points by randomly sampling each star's mass-radius curve and repeat the process to generate a large number of realizations, representing the mass-radius curve supported by the 100 chosen events. Sampling in this manner can form mass-radius curves which violate causality, and thermodynamic constraints. However, we note that this makes our estimates more conservative, and curves which differ

[8] Note that even if Λ_1 and Λ_2 are measured by gravitational-wave observations the tidal Love numbers of the two neutron stars will still be unknown and hence the radii cannot be inferred

greatly from the true EOS as a result of this will be rejected by the chi-square statistic described in the following section. The radii used to construct these mass-radius curves then contain the systematic errors introduced by our use of the universal relations, so the resultant mass-radius curve will also be biased. Given an EOS, we can determine the exact value of this bias by comparing the mass-radius curve for an EOS generated using the TOV equations to that of a curve generated using the universal relations. With this in hand, we can calculate the correction necessary to account for the systematic errors introduced by the universal relations which, when applied to a mass-radius curve, will closely match the exact TOV curve.

In this work, we thus correct for these systematic errors by applying these corrections to the calculated mass-radius curves per EOS. For example, if we would like to determine whether the underlying equation of state of our mass radius curve is ALF2, we first apply the known correction for ALF2 to our mass-radius curves and then complete the comparison described in the next section. If the true underlying EOS is not the one for which we have applied the correction, then the correction will not correctly account for the systematic errors and we can only assume that most similar resulting mass-radius curve is the closest to the excluded true model.

We will consider the true model in turn to be one of the 10 EOS models shown in Fig. 8.1 and show how the corrected-mass-radius curves compare with the true EOS model. In practice, one has to compare the curves with the full set (of millions) of curves. In order to clearly illustrate the power of the method, we have not done so and instead reserved a more detailed and careful Bayesian statistical analysis of model selection in an upcoming publication.

8.5.3 EOS Model Selection Using χ^2 Statistic

After generating a mass radius curve as described in the previous section, we must compare it to a set of EOS models in order to determine the true EOS of the population. We complete this comparison with the following statistic:

$$\chi^2_{k,M} = \frac{1}{N} \sum_{i=1}^{N} \frac{(r_i^k - r_i^M)^2}{\sigma_i^2} \tag{8.19}$$

Here, N is the number of events, k stands for one of the realizations constructed from the mass-radius posterior and σ_i are $1-\sigma$ uncertainty in the radii calculated after applying the systematic bias correction. We generate 500 realizations of the mass-radius curve and obtain a distribution of the χ^2 statistic for each of the 10 EOS models.

If a realization of the mass-radius curve is close to the model to which it is compared to, the numerator of Eq. (8.19) becomes zero. If, however, the uncertainties in the tidal deformability are large, the χ^2 again becomes small regardless of the position of m-Λ posterior distribution with respect to the model m-Λ curve. This is a drawback in our model and leads to the underestimation of near-future LVK upgrades in distinguishing EOS models. Therefore, when comparing against a collection of EOS, the smallest χ^2 value should correspond to the injected EOS for XG detector configurations in which statistical errors are much smaller and recovery of EOS in the data is more accurate, but for near-LVK upgrades, this may not be true due to the large errors in tidal deformability. We discuss our results in the next section and defer the improvement to the Bayesian formulation of our χ^2 method to future studies.

8.5.4 Combining Results from Multiple Events

The accuracy of radii posteriors depends to a large extent on the accuracy of tidal deformability measurements, which in turn depends on mass-posteriors. Heavier component masses have smaller tidal deformabilities, which are difficult to measure. The low accuracy of the tidal deformabilities results in poorer radii measurements, which constrain the high-density regime of the EOS, while lighter component masses typically result in better measurement of the radii. The correct reflection of the radii uncertainty, therefore, cannot be at some fiducial mass but will be a function of the companion mass. Having measured the radii of several hundred neutron stars, it is possible to get a better handle on the radius at a fixed mass.

Evidence from the observation of multiple events, in principle, can be combined to give us integrated evidence of the constraints on neutron star radii. In this paper, we bin the selected set of events over the range of companion masses from 1.0 $\mathrm{M}_{\odot}$ to the maximum mass supported by the EOS in steps of 0.05 $\mathrm{M}_{\odot}$ wide bins and assume that all neutron stars in a given bin have the same radius. The uncertainty in the radius in each bin is computed as the quadratic harmonic sum of individual 1-σ uncertainties in the radius of individual neutron stars that lie within the bin. This procedure is equivalent to combining the posteriors of radii corresponding to all the NSs in a particular mass bin assuming priors are the same for all NSs. While not ideal, this method improves upon the method used in [232], which assumes that radii of neutron stars over the entire mass range from 1.1 $\mathrm{M}_{\odot}$ to 1.6 $\mathrm{M}_{\odot}$ are the same. We note that this latter assumption could introduce an intrinsic systematic error of 200 m (Eq.-6 of [232]) — a value much larger than the measurement uncertainty we find in the case of XG detectors. We report the results of this calculation in the next section. The accuracy of radii measurements can be translated to the accuracy in the estimation of nuclear physics parameters [262, 263] which we defer for future work.

8.5.5 Impacts of assumed cosmology

To obtain the error in the radius measurement, we need to convert the uncertainties in the detector-frame chirp mass to that of the source-frame chirp mass. In doing so, we have assumed that the cosmological parameters, like the Hubble constant (H_0), are known exactly (see Sec. IIA). Although advancements in gravitational-wave detector networks are expected to achieve sub-percent precision in measuring cosmological parameters [264–269], the associated uncertainties may still impact radius measurements.

Note that the two most precise measurements of H_0, from the Planck mission [270] and the SH0ES project [271], are in disagreement at the $5-\sigma$ level, which is called the Hubble tension. To obtain a liberal estimate of how the uncertainty in H_0 can affect radius measurements, we perform Bayesian parameter estimation with `BILBY` [272, 273] for a (1.45, 1.35) $\mathrm{M}_{\odot}$ BNS system, at 400 Mpc, with APR4 as the assumed EOS. For this zero noise analysis, the system is injected in a network with one Einstein Telescope and two Cosmic Explorer observatories (SNR 330). The injected system is made to obey the SH0ES estimate of $H_0 = 73.3\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, whereas the recovery is performed assuming the Planck18 value of $H_0 = 67.4\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, i.e., a fractional error in H_0 of $\sim$ 8%. Employing the same analysis as in the current study, we obtain the 68%-credible region for radius estimate to be 370m ($\Delta R/R$ 3%). In contrast,

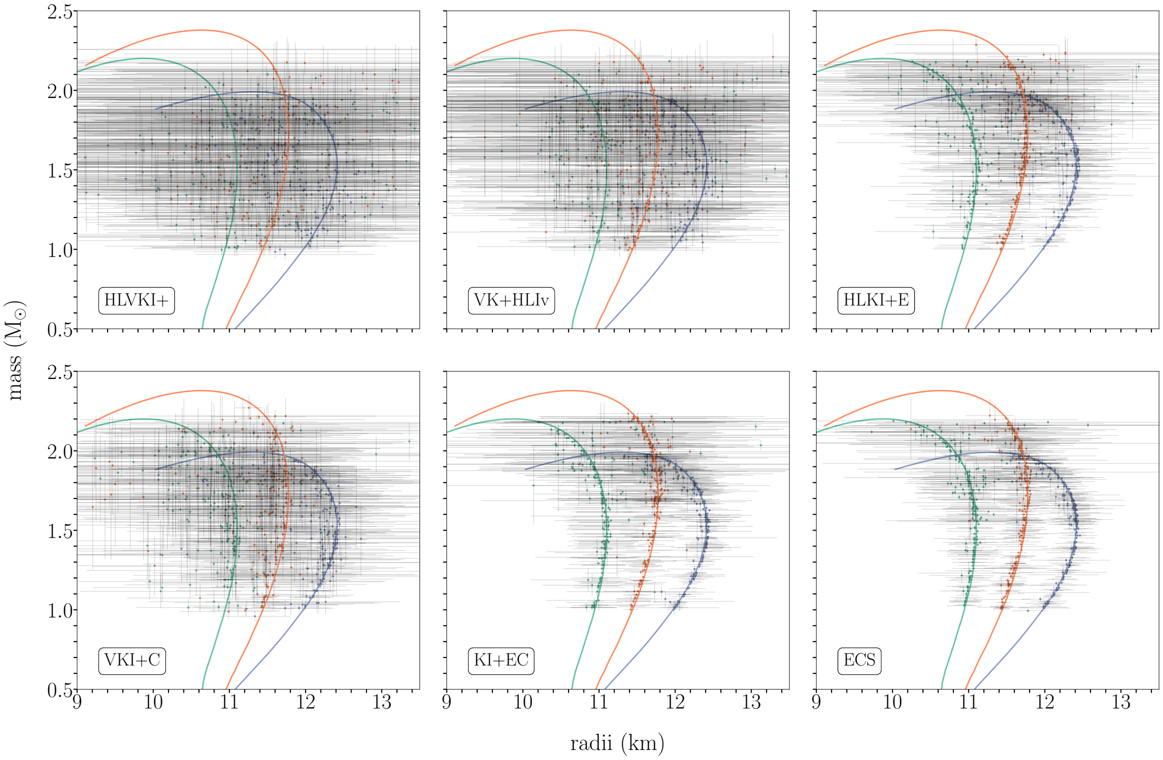


Figure 8.4. TOV mass-radius curves of ALF2(blue), APR3(orange), and APR4(green) overlaid with the bias-corrected recovered mass and radius as well as their errors (grey bars) in a subset of near-future and XG detector networks, for a set of 100 random events drawn from the 500 loudest SNR. There is a clear trend of improving radius error as the detector networks improve left to right, top to bottom. Additionally, in the best detector networks, radius errors also improve with decreasing mass, as is to be expected with higher accuracy in the measurement of higher tidal deformability.

the bias in the estimate due to inference with the incorrect cosmology is 60m. Thus, even at an exaggerated uncertainty of 8% in H_0, we see that the statistical uncertainty in the radius measurement outweighs the resulting bias. Therefore, at the forecasted precision levels of cosmological parameter measurement with next-generation observatories, we do not expect the uncertainty in their estimation to play a significant role in the estimation of the radius of the neutron star.

8.6 Results from a Population Study

In this section, we present the accuracy of radius measurements inferred from a sub-population of 100 best events, for six different detector networks and three different EOS models. The sub-population is chosen to be either events with the best-measured tidal deformabilities or the largest SNRs. In order to gauge Monte Carlo errors, we start with a set of 500 events satisfying the aforementioned criteria and then bootstrap several realizations of 100 events. We present the results in a series of plots that compare the measured mass-radius curves to those

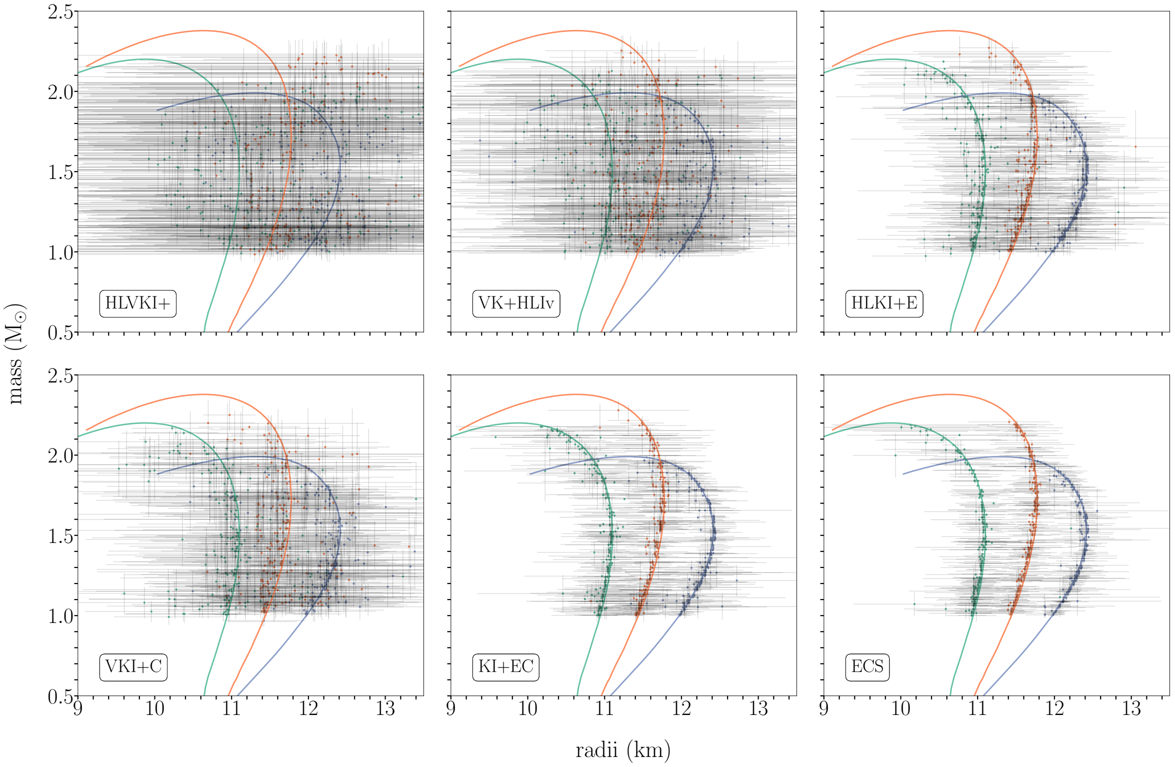


Figure 8.5. Same as Fig. 8.4 except the 100 out of 500 events with best measured tidal deformability are chosen. Again, there is a clear trend of improving radius error as the detectors network improves left to right, top to bottom. Note that the trend of improved radius error with decreased mass is not clear here as it was with the loudest in the SNR set. This is a natural result from the selection of only the best measured combined tidal deformability systems as opposed to those with the best SNR as in Fig 8.4.

derived from different EOS models, the χ^2 histogram between the measured and model radii, and precision with which radius can be measured by combining events in $0.05M_\odot$-wide mass bins for different EOS models.

8.6.1 Radius Measurement

Figures 8.4 and 8.5 plot the uncertainties in the measurement of masses and bias-corrected radii for 100 random events drawn from the 500 events, with the largest SNR and the best-measured tidal deformability, respectively, for a population of BNS described in section 8.2. The cumulative distribution of the measurement uncertainties in the parameters used in this calculation are shown in Fig 8.3. Multiple realizations of the 100 events (out of 500) do not show significant differences in the mass-radius curves and hence we have shown the plots for just one realization. Results are shown for the six different detector networks. In each case, the true model is in turn chosen to be ALF2 (blue), APR3 (orange), or APR4 (green). In these plots, we show the bias-corrected radii using only the injected models as described in 8.5.2. Otherwise, the plot would be too busy; the chi-square plots, to be discussed below, will compare the bias-corrections applied to radii assuming the true EOS model to be any one of the three candidates. Measurement uncertainties in mass and radius are plotted in grey.

Figures 8.10 and 8.11 in Appendix 8.8.1 show the same result but plotted in the chirp mass-symmetric mass ratio space, while Figures 8.12 and 8.13 show the results in the chirp mass-combined tidal deformability space. Figures 8.10 and 8.12 are for events with the best-measured tidal deformability while Figures 8.11 and 8.13 are for events with the largest SNRs. The color shade of the dots in these plots represents the radius uncertainties while the size of the dots is a measure of the SNR of the events as shown in the legend.

Note that these results are based on the Fisher Matrix calculation of the measurement uncertainty. Therefore, the results we see here can be taken as a lower bound of what we might actually expect from a full Bayesian analysis of parameter estimation of these events.

From Figs. 8.4 and 8.5, there is an evident trend of marked improvement in the measurement of the radii as the detector networks themselves improve. The recovered radii fall closer to the injected EOS curve, and the measurement uncertainties vastly decrease as the number of XG observatories in a network rises from 0 to 3. Notably, the maximum uncertainty in the radii, most easily read from the color bars of Figures 8.10-8.13, vary from, in the worst detector, about 2500 m to, in the best network, only about 300 m. At low masses, the disparity is especially clear, and this is a natural result for these networks— particularly the improvement once at least one XG detector added to the network.

It is notable that in networks which contain just one XG detector, the HLKI+E network slightly outperforms that of VKI+C in the measurement of radius error. This is expected for two reasons. First, the HLKI+E network contains one additional detector than that of VKI+C, which inherently improves its sensitivity. Second, the ET sensitivity curve, as seen in Fig. 8.2, contains a long tail in the low-frequency regime not present in the CE curve. This increases the time neutron star signals spend in the band, and results in a better-measured chirp mass and, therefore, better-measured radii. The evidence of this can be seen in the chirp mass and radii CDFs of Figure 8.3. There, the HLKI+E chirp mass CDF shows clearly a smaller relative error than that of VKI+C, and where the HLKI+E tidal deformability CDF shows on level or slightly smaller relative error than that of VKI+C.

In the data set with the loudest SNR events (Figures 8.4, 8.11, and 8.13), higher-mass systems are less constrained—especially in radius—than lower-mass systems, while this is not necessarily true for the set of best-measured tidal deformability events (Figures 8.5, 8.10, 8.12). Again, this is an expected result, as we accumulate most SNR for BNS systems during the low-frequency inspiral phase, while the best measurements of tidal deformability come from the high-frequency part of the waveform during the merger. Thus, a high SNR does not beget a well-measured tidal deformability or radius. Additionally, although gravitational-wave amplitudes for high-mass systems tend to be larger compared to low-mass ones, the value of their tidal deformability tends to be smaller. These small values combined with short inspiral times result in larger relative errors in the measurement of tidal deformability and radii despite the boost in SNR from higher amplitudes. This trend is especially clear in Figs. 8.12 and 8.13 where in Fig. 8.13 the highest radii errors for each panel (in yellow) are always seen in the right, or the high mass and low tidal deformability, portion of the plot while in Fig. 8.12 the worst measured events (again in yellow) are spread throughout parameter space.

Similarly, in Figs. 8.10 and 8.11, it appears that a high symmetric mass ratio, and high chirp masses may result in poorly measured tidal deformability for the highest SNR events, but not necessarily for those with the best measured tidal deformability. In Figure 8.11, large radii errors (in yellow) are typically grouped in the upper right-hand corner of most plots, with a small spread along the right-hand edge in the ALF2 and APR4 EOS, and a small line along the upper-edge in the VKI+C network of ALF2. This is again due to the previously discussed issue with taking the loudest SNR events, but whether this is individually caused by either the high symmetric mass ratio or the large chirp masses is not immediately clear. As previously mentioned, a high chirp mass comes with a small tidal deformability and therefore large relative error. However, a high symmetric mass ratio can also decrease the inspiral time, or time in a frequency band, and therefore again the accuracy of the measurements becomes low. Notably, in the set of best measured tidal deformability shown in Fig. 8.10, the large errors are distributed more evenly throughout the plot and have lower maximums than their high SNR equivalent.

8.6.2 Model Selection

Figures 8.6, and 8.7 show the primary results from our model selection procedure. Here we plot the distribution of the χ^2 statistic defined in Eq. (8.19) between the observed mass-radius curve and the one predicted by the chosen EOS model. The separation of the distribution for any two EOS signifies the effectiveness of a detector in distinguishing between the injected and test EOS models. In these figures, each row corresponds to a particular detector network, while each column corresponds to a specific injected EOS (label at the top of the column). The χ^2 histograms in each panel are additionally colored to match the EOS color scheme as in Fig. 8.1, with the count on the y-axis and the χ^2 (in log-scale) on the x-axis.

For detector networks in the top two rows the inferred radius r^k is very different from that predicted radius r^M by any of the models (see top left and middle panels in Figs. 8.4 and 8.5) which would cause χ^2 to be large. However, at the same time, the uncertainties in the measurement (σ_i) are also large. Consequently, for networks with poorer sensitivity, the χ^2 will tend to be equally small no matter which EOS model the events are compared to.

The story is different when the radius uncertainty σ of a detector network is small. For such detectors, the bias-corrected radius differs significantly from the predicted radius found using a model other than the true one, but agrees very well with the predicted radius of the true model. Consequently, the ratio within the sum in Eq. (8.19) is small only when the set $\{r_i^M\}$ corresponds to the true EOS. This is the reason why the χ^2 distributions for the models other than the true one have far greater values than they are for the true model in the bottom two rows. We find that the method accurately recovers the injected EOS model among a larger set of models than was used for injection. In addition, we have also used a much larger sample of events for our work compared to previous studies [239, 274–276].

We stress that the power of the χ^2 statistic introduced lies in discriminating between the different EOS models when measurement uncertainties are small; with less sensitive detector networks there is no way to distinguish one EOS model from another. The absolute value of the χ^2, however, has no significance.

Across different detector networks, when the injected EOS is close in the M-R parameter space to the comparison EOS, the distribution is most often confused with the true EOS as

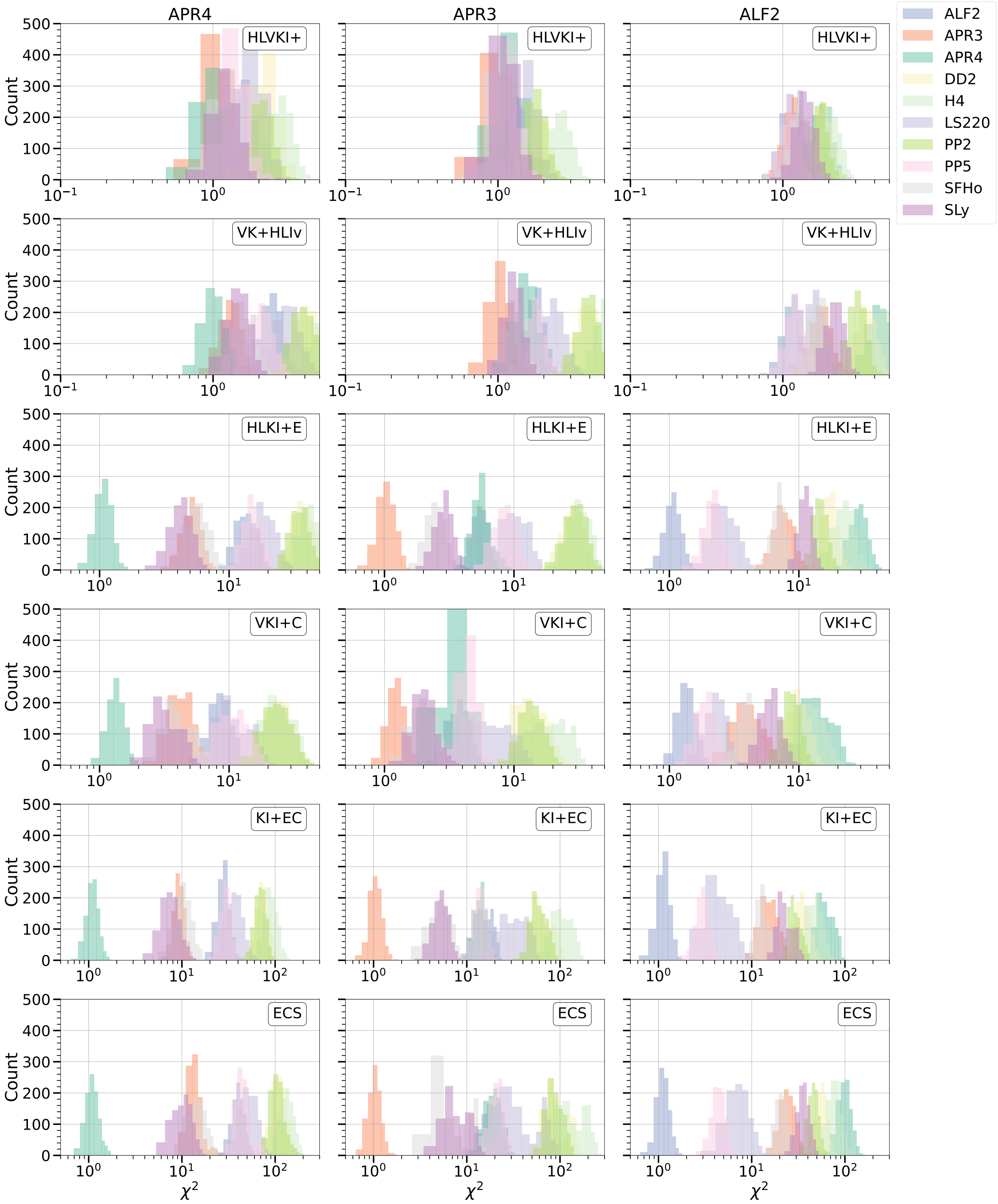


Figure 8.6. Chi-square histograms for 100 events from those 500 with the smallest error in combined tidal deformability. The injected EOS is listed along the top, and the colored histograms represent the result assuming a second EOS model, including the original injection. Detector networks are organized by sensitivity row-wise with the most sensitive network at the bottom. In every EOS and network scenario including at least one XG detector, the injected EOS is recovered correctly and easily distinguishable from the other nine via this test. In our two least sensitive and nearest future detector networks, HLVKI+ and VK+HLIvc, the opposite is clearly true and all models are indistinguishable.

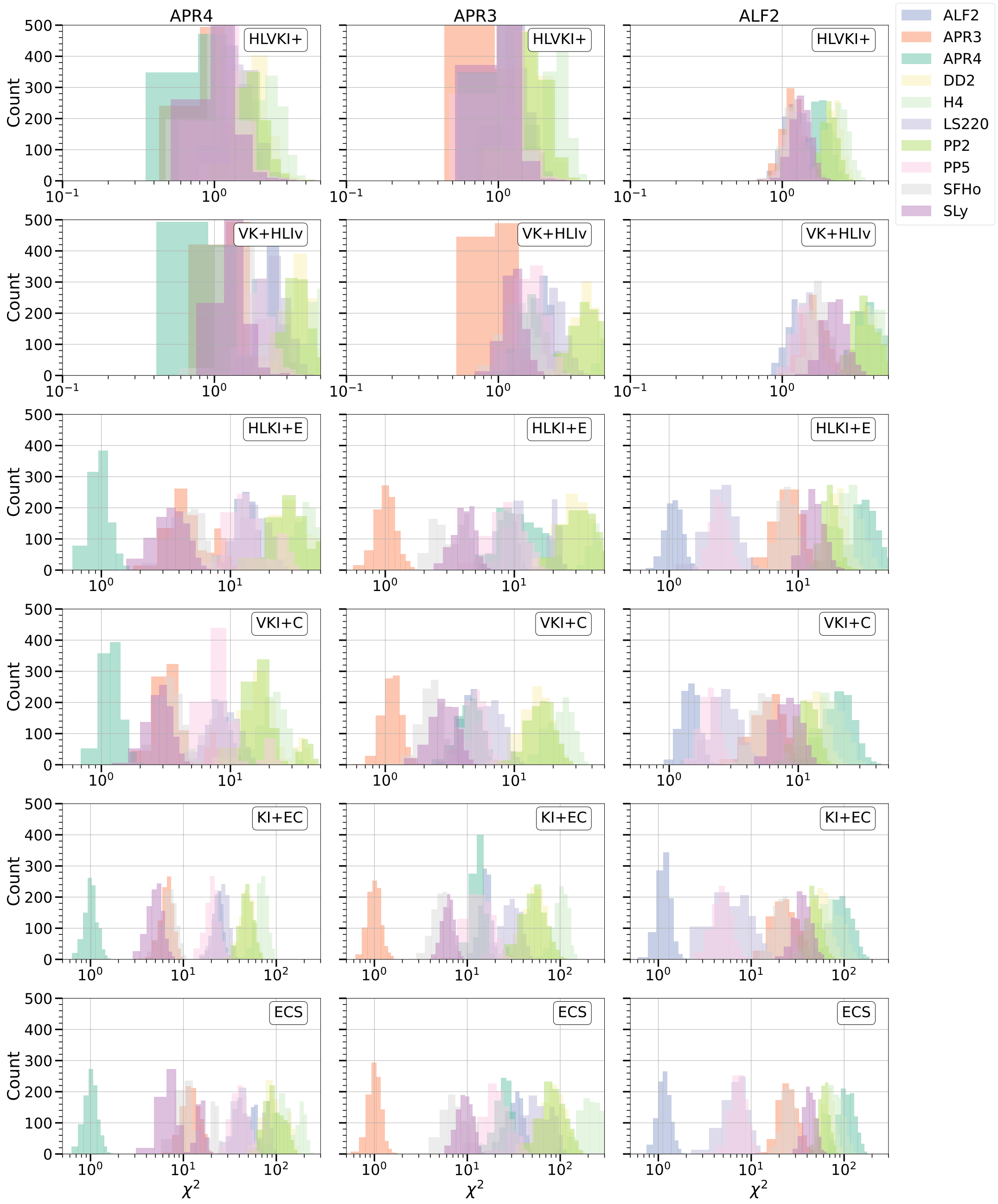


Figure 8.7. Chi-square histograms for 100 events from those 500 with the smallest error in combined tidal deformability. The injected EOS is listed along the top, and the colored histograms represent the result assuming a second EOS model, including the original injection. Detector networks are organized by sensitivity row-wise with the most sensitive network at the bottom. While the peak of the injected EOS histogram is generally recovered with the smallest χ^2 value despite detector sensitivity, in the two networks which do not contain at least one XG detector, the histograms are not distinguishable and we cannot claim that this test is effective in distinguishing EOSs at loud SNRs. However, in networks with at least one XG detector, the correct EOS is consistently recovered with its distribution clearly separated from other EOS models. The same trend is also seen in Fig 8.6.

show by the proximity of its histogram to the true one. For example, in the least sensitive detector networks, or top rows of Fig 8.7 and 8.6, the overlap between the resultant three distributions of ALF2, SLy, and PP5 is total, and even with one XG detector, they still overlap significantly. It is only in the best detector networks (bottom two rows) that they begin to become indistinguishable. Meanwhile, comparing ALF2 to APR4 or H4, even in some of the least sensitive networks, their distributions already diverge from the true ALF2 one. This follow from the simple fact that at low mass in the mass-radius curve, ALF2 lies very close to PP5 and at high mass close to SLy and would therefore naturally match more closely with its nearest neighbors while the distance between ALF2 and APR4 or H4 is significant and therefore not well matched (Kashyap et al. [261] discuss how distinguishability of EOS models changes with respect to the L_2 distance between them).

In general, as the sensitivity of the networks increases, so too does the separation of the posterior distributions. In the lower sensitivity networks from both the highest SNR and best measured tidal deformability data sets, the distributions overlap significantly, and it is only with the inclusion of at least one XG detector that the distributions become at all distinguishable. Across EOS and data sets in networks with at least one XG detector, the smallest χ^2 value always corresponds to the injected EOS and its peak is distinguishable from the EOS with the next smallest χ^2 value. There is not a significant separation of the true EOS from its neighbors, however, until we begin to include at least two XG detectors in the network. In these most sensitive networks, the true EOS centers around one, effectively recovering the EOS, and there are an order of one hundred separations between it and its neighbors, giving hope that XG detector networks may be able to distinguish clearly between these, and other EOSs.

8.6.3 Combining Radii Errors from Multiple Events

In Fig. 8.8, we present the results of combining the radius uncertainties of multiple events binned in individual masses of neutron stars in the range 1.0 $M_\odot$ to the maximum mass supported by the EOS used for the injection, using the method described in Sec. 8.5.4. We plot the effective errors in the radii of a particular mass bin for three EOS models, with the colors the same as in Fig. 8.1. The color bands show the variation in the combined error due to bootstrapping while selecting 100 events out of 500 best events according to two different criteria (best SNR and best measured $\tilde{\Lambda}$) as described in the previous sections. We've found that this selection of events does not make a significant difference to the results.

One of the crucial features of these plots is the increase in the effective radius uncertainty with the increase in the masses of the individual neutron stars. This is again due to the small tidal deformability of heavy neutron stars and poor accuracy in their measurements irrespective of the EOS and the detector network, leading to the poor measurement of radii via the C-Λ universal relation. Smaller radii and tidal deformabilities at higher masses result in poorer constraints of the EOS at higher densities, which is usually near the neutron star core. As expected, we find an improvement in the radii uncertainties for all mass bins according to the $\sqrt{N}$ law, where N is the number of events combined in each mass bin.

We find the uncertainties to be smaller than 1 km by combining 5 or more events in any mass bins irrespective of the network chosen. The HLVKI+ network has typical errors to be around 1 km for all masses and becomes as large as 3 km even after combining multiple events. The addition of one XG detector to the network improves the radii uncertainties by an order-of-magnitude with a typical value of 100 m, the smallest value of 30 m, and the largest value of 1 km. The best radius measurement, however, is accomplished by combining both ET and CE. We show the results for two such networks of detectors where uncertainties could be as small as 20 m with almost all of the bins having uncertainties smaller than 100 m (i.e., ~ 1 %). We emphasize again that in these calculations, we use Fisher Information Matrix to approximate the uncertainties, which are a lower bound. We defer the work of accurate analysis using Bayesian Monte Carlo methods to future work.

8.6.4 Discussion

The result of our analysis for the best-SNR and best-measured tidal deformability data sets is promising for networks including XG observatories. Advanced LIGO and Virgo and their upgrades in the near future are expected to observe tens of events with moderate SNR (i.e., SNR> 40) and a handful of high-fidelity (SNR> 100) events over a two-year period (cf. 8.2, columns 2 and 3). Without any XG observatories, the best fractional uncertainty in radii measurements for the top 100 events with best measured tidal deformability is 5–10%, with more than half above 10%, as seen in Fig 8.9. This means it will be difficult for these networks to distinguish between even the most disparate set of EOS models considered in this paper. However, with the inclusion of just one XG detector, the best results show only a 0.8% uncertainty in radii, with half of the events reporting only 6% or less, allowing EOS to become partially distinguishable. Meanwhile, networks with at least two XG detectors tell a completely different story.

In our most sensitive networks, we will be able to measure the radii of neutron star sources to 0.5%, with half at 3% or less, as seen in Fig 8.9. However, we have not taken into account the models of the crust of neutron stars which themselves can be 100 m (i.e., 1% of the radius), so further work is required to better characterize the meaning of measurement accuracies below this accuracy. These precise measurements, however, result in χ^2 distributions that are easily distinguishable, well separated, and centered for both the loudest SNR and best-measured tidal deformability event sets. Consequently, XG networks will be able to distinguish between different EOS models (even ones that are sufficiently close to each other in L_2 measure of distance) and place stringent constraints. Overall, the results of these data sets reveal an avenue for future research that deserves to be pursued further.

8.7 Summary and Conclusions

In this work, we report on the improvements in the inference of the dense matter equation of the state of neutron stars with the current and next-generation gravitational-wave detectors based on their expected design sensitivity curves. We evaluate the measurement uncertainties for hundreds of thousands of events and consequently, it is not possible to carry out a Bayesian inference analysis of the events as that would currently take a formidable amount of time. Instead, we use the Fisher matrix approximation to compute the 1-σ uncertainties and correlations of the binary neutron star parameters, including the companion masses and the effective tidal deformability $\tilde{\Lambda}$ using the IMRPhenomPv2NRTidal waveform. The multivariate distributions of the binary parameters obtained from gravitational-wave observations, together with two universal relations, namely, Eqs.(8.16) and (8.18), allow us to infer the mass-radius posteriors of companion neutron stars. Since the universal relations are not exact, the inferred radii posteriors have systematic biases. We have shown that these systematic biases can be corrected for when comparing the measured mass-radius posteriors with that predicted by a specific equation-of-state model. Our bias-correction method is equivalent to comparing the model mass-tidal deformability predictions directly with the gravitational-wave data but computationally inexpensive since bias corrections are known a priori and don't need to be generated on the fly. Moreover, the method avoids having to repeat the likelihood calculations and computations of posteriors for every plausible equation-of-state.

We employed this new method to compare three disparate model equations of state with simulated gravitational-wave measurements for assuming the true equation of state to be one of the 10 models. Our results demonstrate that the method can uniquely identify the correct equation-of-state when the detector network contains at least one XG observatory (either Einstein Telescope or Cosmic Explorer). It will be difficult to distinguish between different plausible equations of state with the current network of LIGO, Virgo and KAGRA observatories or their proposed improvements (A+ or Voyager). However, with the addition of at least one XG observatory, it will be possible to draw firm conclusions about the true equation-of-state describing dense matter in neutron star cores. Moreover, we find vast improvements in the measurement uncertainties of neutron star radii with two or more next-generation observatories in the network. More specifically, we find that radius uncertainties are a few hundred meters for networks with one or more next-generation observatories, while this would be 1 km in a network with the LIGO-Virgo-KAGRA network and their future upgrades. However, we found that the overall accuracy of radii measurements decreases with increasing neutron star mass. This is because tidal deformabilities are smaller and more difficult to measure for more massive neutron stars.

Building more sensitive gravitational-wave observatories is crucial to constraining plausible EOS models—measurements that can inform not only the gravitational-wave community but also the nuclear physics and astronomy communities at large. In this light, the radius of a typical NS can be constrained to better than 30 m, at the lower end of the expected range of neutron star masses, with joint detections of events over two years in Einstein Telescope and Cosmic Explorer.

8.8 Appendix

8.8.1 Miscellaneous plots

In this section we assemble a list of four additional plots to gain a better understanding of the results presented in the main body of this paper. These plots show the measurement uncertainty in radius either as a function of chirp mass and symmetric mass ratio in Figs. 8.11 (for 100 randomly chosen events out of the 500 loudest events) and 8.10 (for 100 randomly chosen events out of 500 events with the best measured tidal deformability) or as a function of the chirp mass and tidal deformability in Figs. 8.13 (for 100 randomly chosen events out of the 500 loudest events) and 8.12 (for 100 randomly chosen events out of 500 events with the best measured tidal deformability).

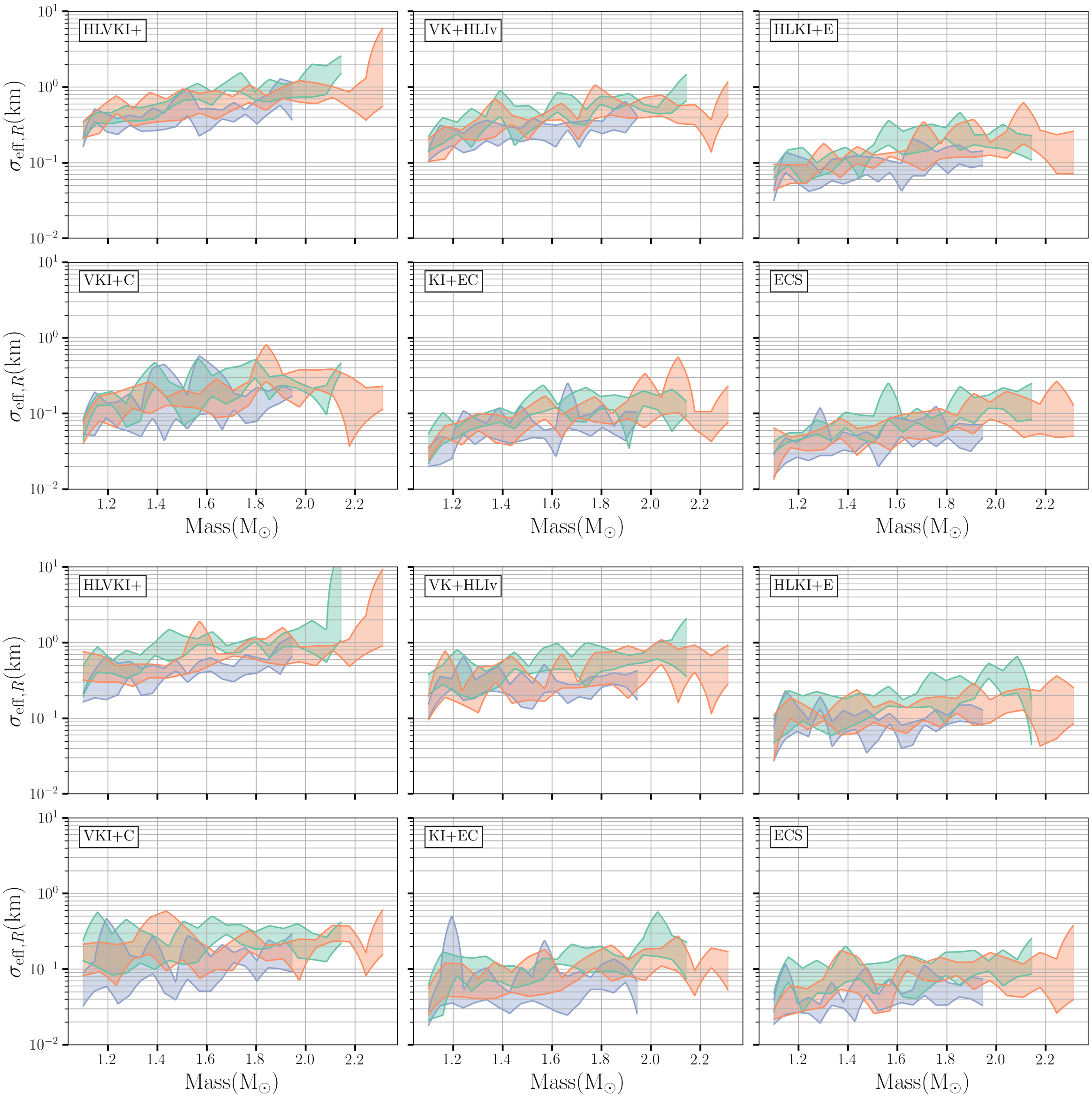


Figure 8.8. Cumulative radius error in each mass bin by square harmonic sum assuming constant radii in each mass bin. The *upper panel* shows the 100 events randomly selected from the 500 events with the best measurement of $\tilde{\Lambda}$ while the *bottom panel* shows the same result for 100 events randomly selected from 500 events with the best SNR. The band for each EOS shows the uncertainty due to random sampling. The color encodes the results for each EOS and is the same as Fig. 8.1. We find generically that errors in radii are larger for larger masses across detector networks and data sets due to smaller accuracy in the measurement of smaller tidal deformability.

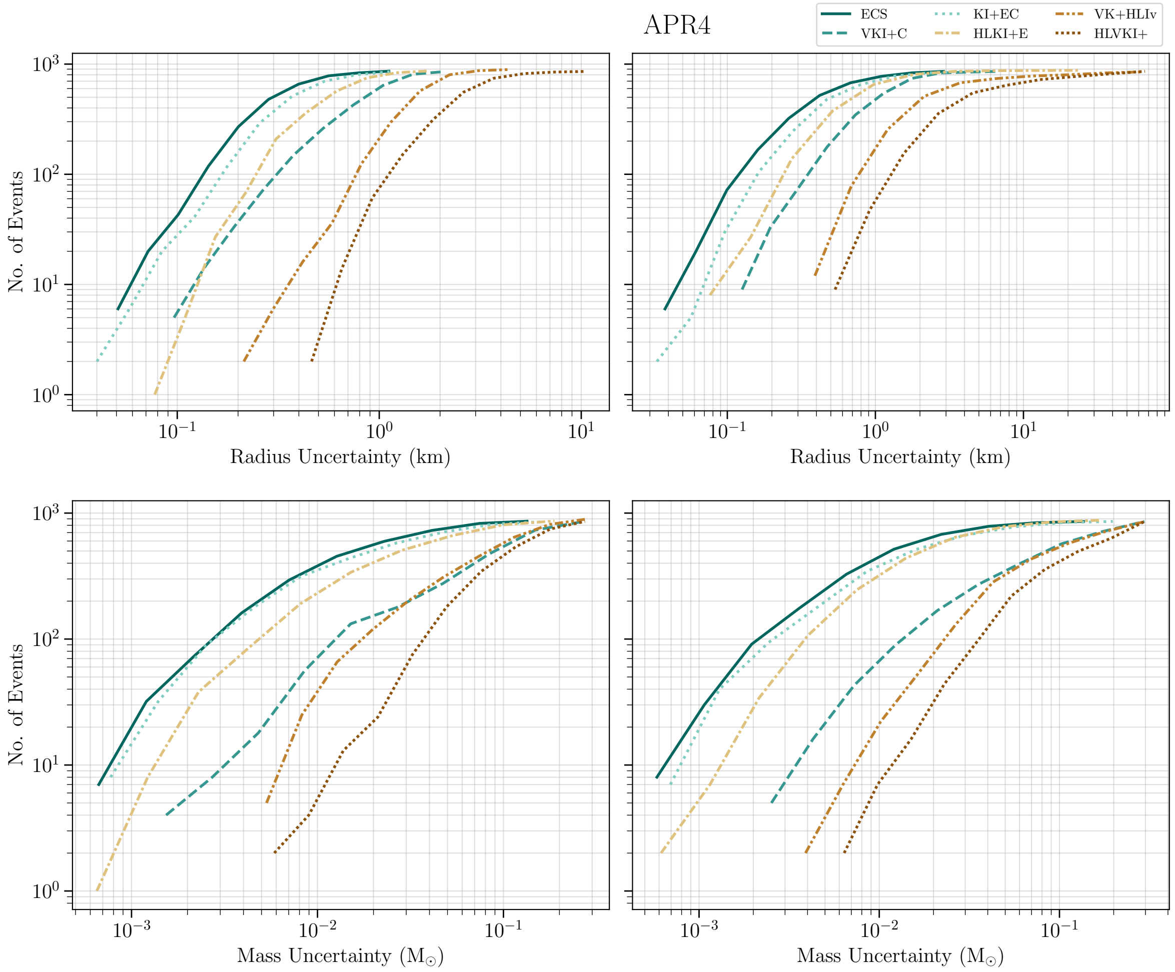


Figure 8.9. *Upper Panel:* Cumulative histograms of the uncertainty in neutron star radii in km (top two panels) and masses in solar mass (bottom two panels) multiplied by the total number (860) of neutron stars in the 430 selected BNS events. The left panels are for events with the best measured tidal deformability and the right panels are for events with the highest SNRs. The different curves correspond to different detector networks considered in this study. These plots show that even the inclusion of just one XG detector (VKI+C or HLKI+E) leads to a vast improvement in the precision of radii measurements. Such detectors could measure the radius to within about 200 m for several events. A network containing two or three XG detectors would improve by a factor of a few. On the other hand, companion masses are better measured by a network that has ET (0.01 $M_\odot$ to 0.001 $M_\odot$) whose lower frequency performance helps in more accurate determination of the chirp mass and the mass ratio.

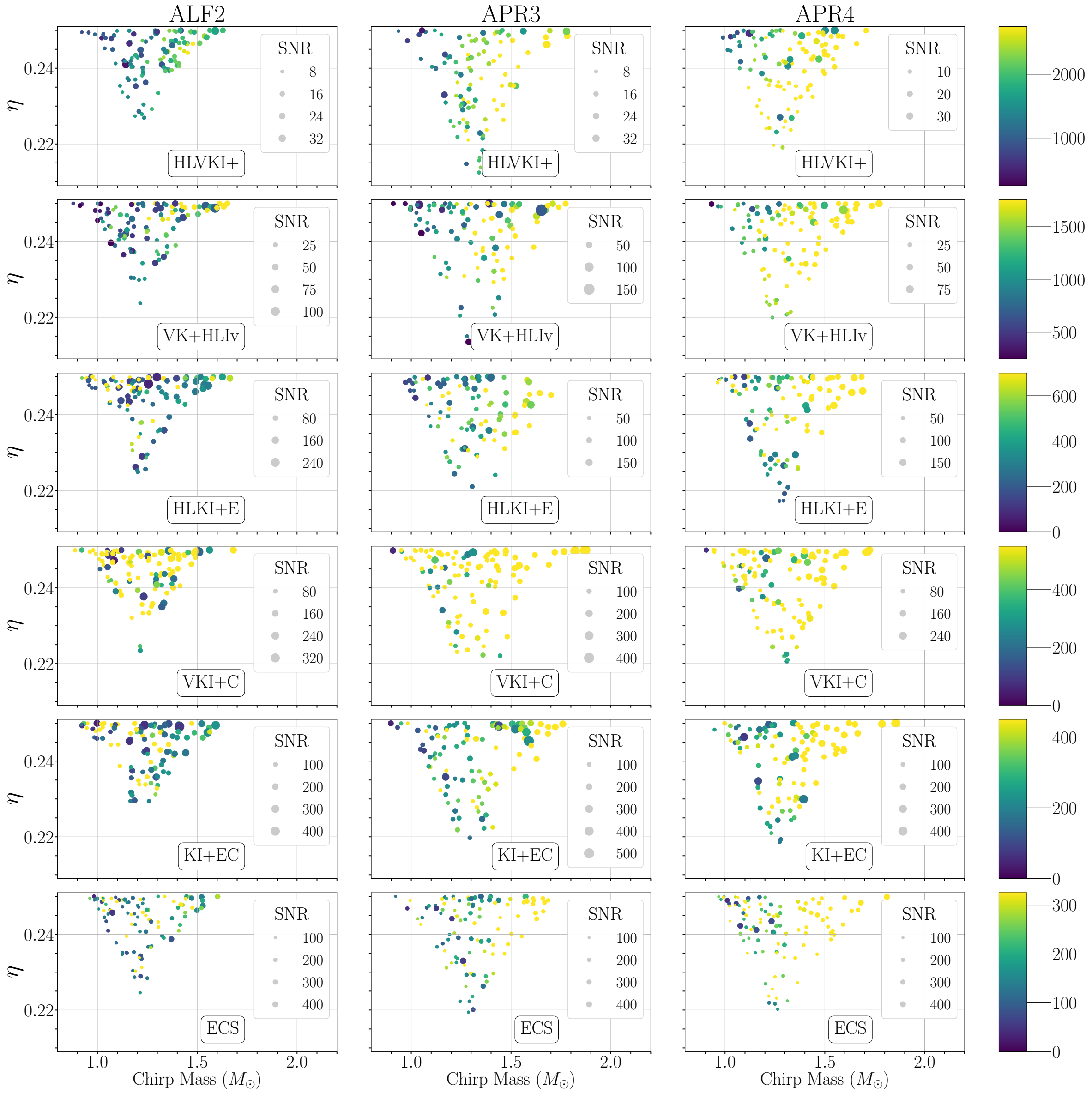


Figure 8.10. The plot shows the radius error (in color) for 100 events with the smallest error in the combined tidal deformability as a function of the symmetric mass ratio and chirp mass. Results are shown for the six detector networks (labelled in each panel) and and three different EOSs (labelled at the top of each column). We recover an approximate trend of increasing radii error with increasing chirp mass and symmetric mass ratio.

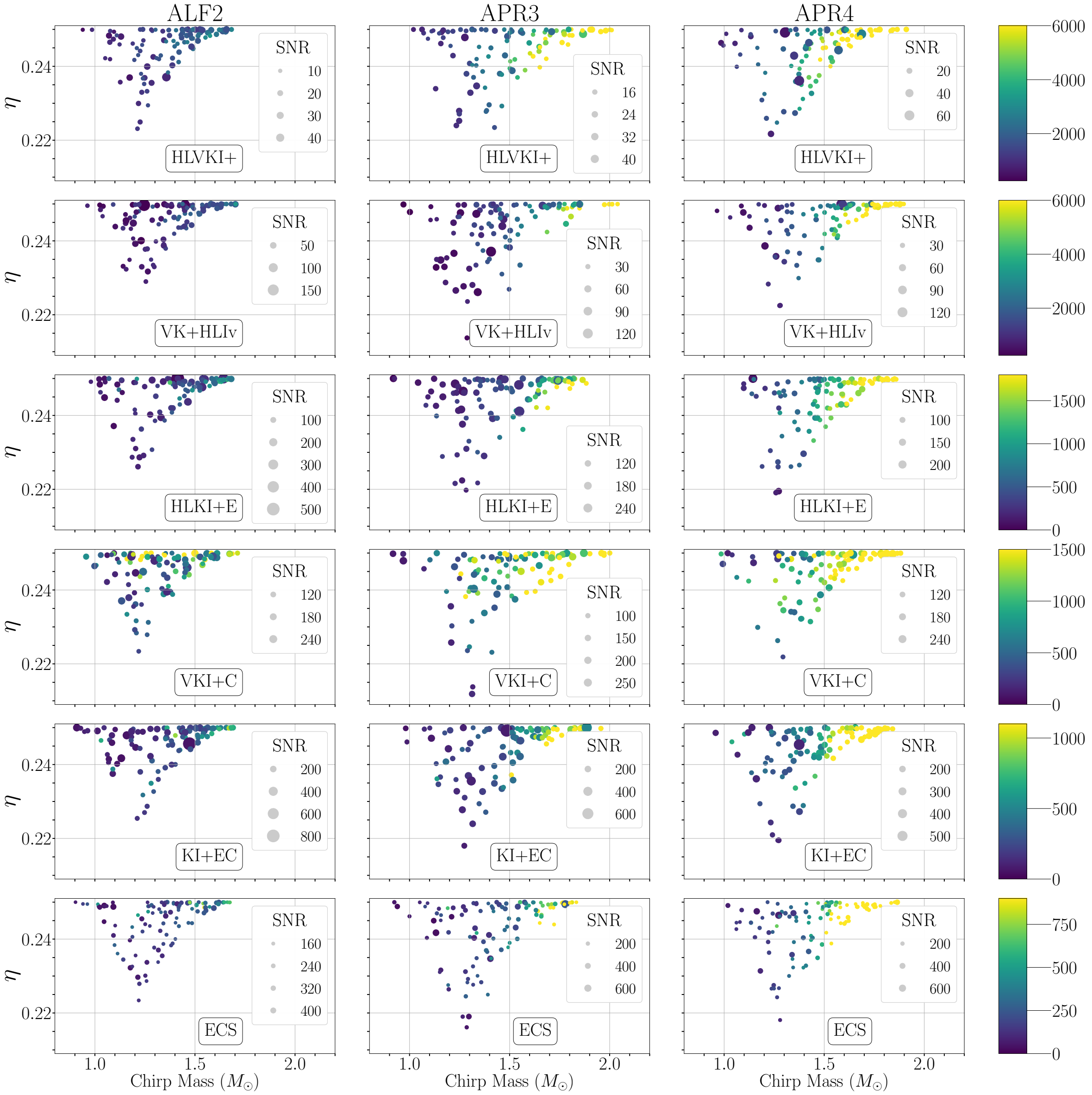


Figure 8.11. Radius error (in color) for 100 random events out of the 500 which are loudest in SNR for our six 3G detectors and three EOS of choice: left ALF2, middle APR3, right APR4. Detectors are ordered top to bottom as follows: ESa4cCa4c, KI+ECa4c, VKI+Ca4c, HLKI+E, VK+HLIvc, HLVKI+. We see the same trend of increasing radii error with increasing chirp mass and symmetric mass ratio as seen in Fig 8.10, but it appears more clearly in this data set, especially in increasing chirp mass.

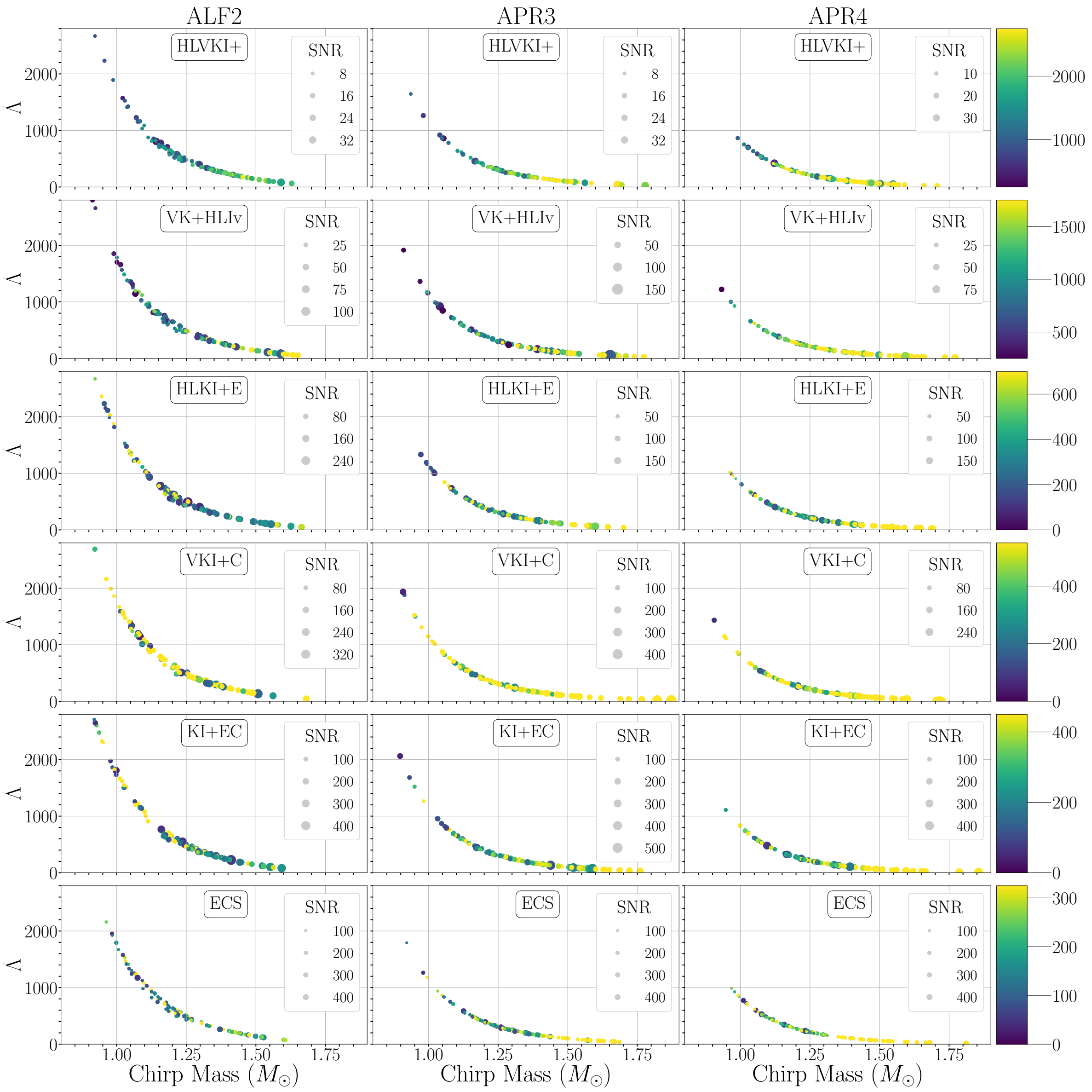


Figure 8.12. Radius errors (in m) are shown in color for 100 random events out of the 500 which are best measured in tidal deformability for our six 3G detectors and three EOS of choice: left ALF2, middle APR3, right APR4. Detectors are ordered top to bottom as follows: ESa4cCa4c, KI+ECa4c, VKI+Ca4c, HLKI+E, VK+HLIvc, HLVKI+. The relative error in radius naturally decreases for systems with larger combined tidal deformability, and again we see larger radii errors at higher chirp mass. Additionally, we naturally see the largest radii errors and smallest SNRs in near-future detector networks, and significantly better ones in networks with XG detectors.

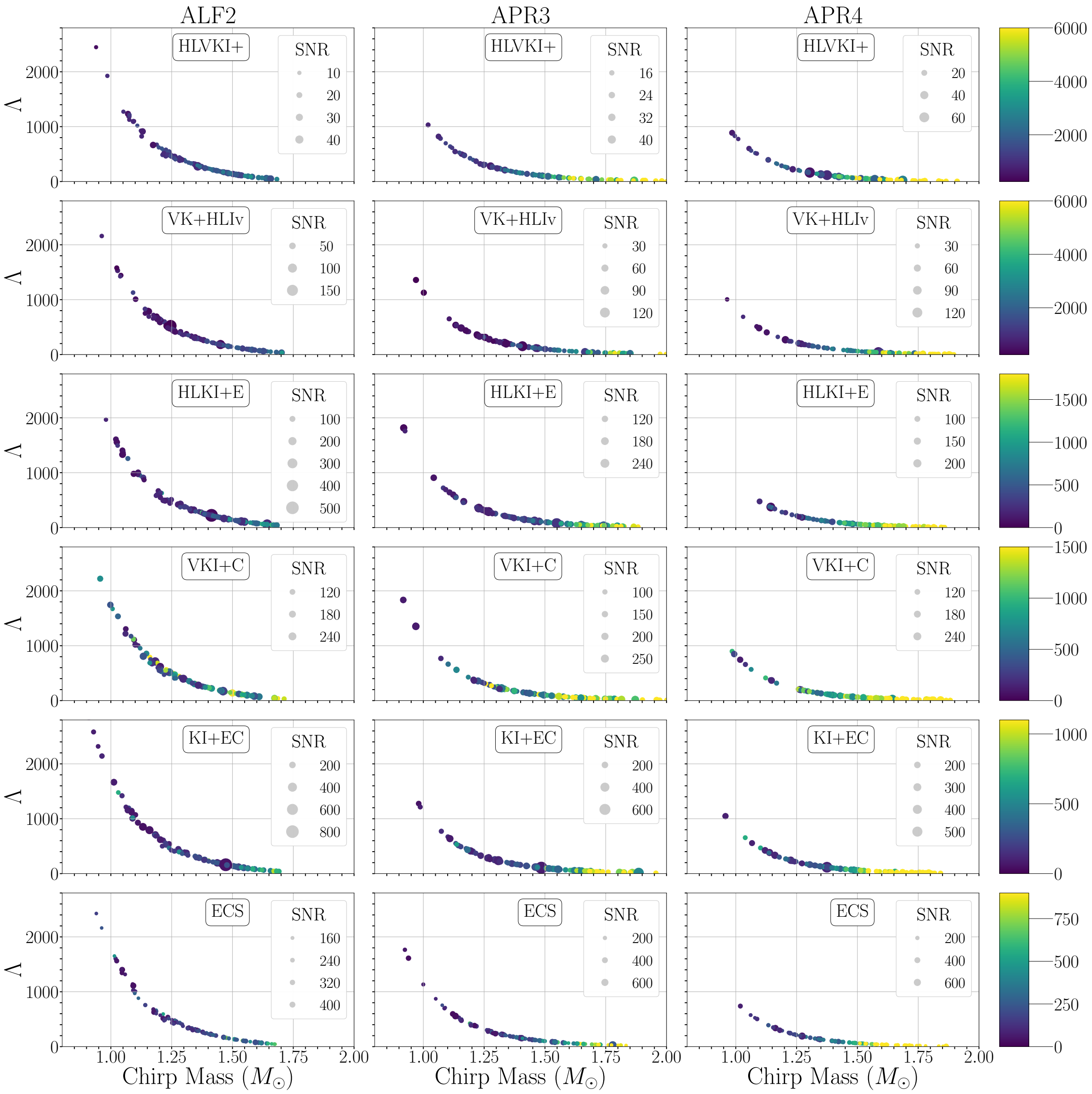


Figure 8.13. Radius error (in color) for 100 random events out of the 500 which are loudest in SNR for our six 3G detectors and three EOS of choice: left ALF2, middle APR3, right APR4. Detectors are ordered top to bottom as follows: ESa4cCa4c, KI+ECa4c, VKI+Ca4c, HLKI+E, VK+HLIvc, HLVKI+. The same trends appear here as in Fig 8.12. Radius error increases with chrip mass, and decreases with combined tidal deformability. Additionally, as the detector networks themselves improve, we see clear improvements radius error and SNR.

Bibliography

[1] Sathyaprakash, B. S. and B. F. Schutz (2009) “Physics, Astrophysics and Cosmology with Gravitational Waves,” *Living Rev. Rel.*, **12**, p. 2, 0903.0338.

[2] “LIGO’s Interferometer,” [Online; accessed April 1, 2024].
URL https://www.ligo.caltech.edu/page/ligos-ifo

[3] et al (LIGO Scientific), J. A. (2015) “Advanced LIGO,” *Classical and Quantum Gravity*, 2312.01211.

[4] Bahaadini, S., V. Noroozi, N. Rohani, S. Coughlin, M. Zevin, J. R. Smith, V. Kalogera, and A. Katsaggelos (2018) “Machine learning for Gravity Spy: Glitch classification and dataset,” *Info. Sci.*, **444**, pp. 172–186.

[5] Sakon, S. et al. (2024) “Template bank for compact binary mergers in the fourth observing run of Advanced LIGO, Advanced Virgo, and KAGRA,” *Phys. Rev. D*, **109**(4), p. 044066, 2211.16674.

[6] Abbott, R. et al. (2023) “Population of Merging Compact Binaries Inferred Using Gravitational Waves through GWTC-3,” *Phys. Rev. X*, **13**(1), p. 011048, 2111.03634.

[7] Madau, P. and M. Dickinson (2014) “Cosmic Star Formation History,” *Ann. Rev. Astron. Astrophys.*, **52**, pp. 415–486, 1403.0007.

[8] S. Soni, D. D. F. D. A. E. T. F. J. G. E. G. G. G. A. H.-C. B. H. B. M. G. M. D. N. A. N. S. N. K. P. A. I. R. R. M. S. S. A. S. M. T., B. Berger (2024) “LIGO Detector Characterization in the First Half of the Fourth Observing Run,” .

[9] Ewing, B. et al. (2024) “Performance of the low-latency GstLAL inspiral search towards LIGO, Virgo, and KAGRA’s fourth observing run,” *Phys. Rev. D*, **109**(4), p. 042008, 2305.05625.

[10] Essick, R., L. Blackburn, and E. Katsavounidis (2013) “Optimizing Vetoes for Gravitational-Wave Transient Searches,” *Class. Quant. Grav.*, **30**, p. 155010, 1303.7159.

[11] Essick, R., P. Godwin, C. Hanna, L. Blackburn, and E. Katsavounidis (2020) “iDQ: Statistical Inference of Non-Gaussian Noise with Auxiliary Degrees of Freedom in Gravitational-Wave Detectors,” 2005.12761.

[12] Abbott, B. P. et al. (2018) "GW170817: Measurements of neutron star radii and equation of state," *Phys. Rev. Lett.*, **121**(16), p. 161101, 1805.11581.

[13] Abbott, R. et al. (2023) "GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo during the Second Part of the Third Observing Run," *Phys. Rev. X*, **13**(4), p. 041039, 2111.03606.

[14] B. Moe, B. S. E. K. R. W., P. Brady and F. Zhang (2014), "GraceDB: A Gravitational Wave Candidate Event Database," .
URL https://dcc.ligo.org/LIGO-T1400365

[15] (2024) "Observation of Gravitational Waves from the Coalescence of a 2.5 − 4.5 $M_\odot$ Compact Object and a Neutron Star," 2404.04248.

[16] Branchesi, M. (2016) "Multi-messenger astronomy: gravitational waves, neutrinos, photons, and cosmic rays," *J. Phys. Conf. Ser.*, **718**(2), p. 022004.

[17] Abbott, B. P. et al. (2017) "GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral," *Phys. Rev. Lett.*, **119**(16), p. 161101, 1710.05832.

[18] ——— (2017) "Multi-messenger Observations of a Binary Neutron Star Merger," *Astrophys. J. Lett.*, **848**(2), p. L12, 1710.05833.

[19] ——— (2016) "Observation of Gravitational Waves from a Binary Black Hole Merger," *Phys. Rev. Lett.*, **116**(6), p. 061102, 1602.03837.

[20] Morisaki, S. and V. Raymond (2020) "Rapid Parameter Estimation of Gravitational Waves from Binary Neutron Star Coalescence using Focused Reduced Order Quadrature," *Phys. Rev. D*, **102**(10), p. 104020, 2007.09108.

[21] Creighton, J. D. E. and W. G. Anderson (2011) *Gravitational-wave physics and astronomy: An introduction to theory, experiment and data analysis*.

[22] Saulson, P. R. (2017) *Fundamentals of Interferometric Gravitational Wave Detectors*, 2nd. ed. ed., World Scientific.

[23] Messick, C. et al. (2017) "Analysis Framework for the Prompt Discovery of Compact Binary Mergers in Gravitational-wave Data," *Phys. Rev. D*, **95**(4), p. 042001, 1604.04324.

[24] Davis, D. et al. (2021) "LIGO detector characterization in the second and third observing runs," *Class. Quant. Grav.*, **38**(13), p. 135014, 2101.11673.

[25] Glanzer, J. et al. (2023) "Data quality up to the third observing run of advanced LIGO: Gravity Spy glitch classifications," *Class. Quant. Grav.*, **40**(6), p. 065004, 2208.12849.

[26] Mukund, N., S. Abraham, S. Kandhasamy, S. Mitra, and N. S. Philip (2017) "Transient Classification in LIGO data using Difference Boosting Neural Network," *Phys. Rev. D*, **95**(10), p. 104059, 1609.07259.

[27] Wu, Y., M. Zevin, C. P. L. Berry, K. Crowston, C. Østerlund, Z. Doctor, S. Banagiri, C. B. Jackson, V. Kalogera, and A. K. Katsaggelos (2024) "Advancing Glitch Classification in Gravity Spy: Multi-view Fusion with Attention-based Machine Learning for Advanced LIGO's Fourth Observing Run," 2401.12913.

[28] Zevin, M. et al. (2017) "Gravity Spy: Integrating Advanced LIGO Detector Characterization, Machine Learning, and Citizen Science," *Class. Quant. Grav.*, **34**(6), p. 064003, 1611.04596.

[29] Robinet, F., N. Arnaud, N. Leroy, A. Lundgren, D. Macleod, and J. McIver (2020) "Omicron: a tool to characterize transient noise in gravitational-wave detectors," *SoftwareX*, **12**, p. 100620, 2007.11374.

[30] McIver, J. L. (2015) *The impact of terrestrial noise on the detectability and reconstruction of gravitational wave signals from core-collapse supernovae*, Ph.D. thesis, Massachusetts U., Amherst, Massachusetts U., Amherst.

[31] Godwin, P. (2020) *Low-latency Statistical Data Quality in the Era of Multi-Messenger Astronomy*, Ph.D. thesis, Penn State U.

[32] Lopez, M., V. Boudart, S. Schmidt, and S. Caudill (2022) "Simulating Transient Noise Bursts in LIGO with gengli," 2205.09204.

[33] Vallisneri, M., J. Kanner, R. Williams, A. Weinstein, and B. Stephens (2015) "The LIGO Open Science Center," *J. Phys. Conf. Ser.*, **610**(1), p. 012021, 1410.4839.

[34] Adhikari, R. X., A. Brooks, B. Shapiro, D. Mcclelland, E. K. Gustafson, V. Mitrofanov, K. Arai, C. Wipf, and H. Ligo (2023) *LIGO Voyager Upgrade: Design Concept*, *Tech. Rep. LIGO-T1400226-v9*, LIGO SCIENTIFIC COLLABORATION.
URL https://docs.ligo.org/voyager/voyagerwhitepaper/main.pdf

[35] Saleem, M. et al. (2022) "The science case for LIGO-India," *Class. Quant. Grav.*, **39**(2), p. 025004, 2105.01716.

[36] Akutsu, T. et al. (2019) "KAGRA: 2.5 Generation Interferometric Gravitational Wave Detector," *Nature Astron.*, **3**(1), pp. 35–40, 1811.08079.

[37] Punturo, M. et al. (2010) "The third generation of gravitational wave observatories and their science reach," *Class. Quant. Grav.*, **27**, p. 084007.

[38] Sathyaprakash, B. et al. (2012) "Scientific Objectives of Einstein Telescope," *Class. Quant. Grav.*, **29**, p. 124013, [Erratum: Class.Quant.Grav. 30, 079501 (2013)], 1206.0331.

[39] Maggiore, M., C. V. D. Broeck, N. Bartolo, E. Belgacem, D. Bertacca, M. A. Bizouard, M. Branchesi, S. Clesse, S. Foffa, J. García-Bellido, S. Grimm, J. Harms, T. Hinderer, S. Matarrese, C. Palomba, M. Peloso, A. Ricciardone, and M. Sakellariadou (2020)

"Science case for the Einstein telescope," *Journal of Cosmology and Astroparticle Physics*, **2020**(03), p. 050.
URL `https://dx.doi.org/10.1088/1475-7516/2020/03/050`

[40] Reitze, D., R. X. Adhikari, S. Ballmer, B. Barish, L. Barsotti, G. Billingsley, D. A. Brown, Y. Chen, D. Coyne, R. Eisenstein, M. Evans, P. Fritschel, E. D. Hall, A. Lazzarini, G. Lovelace, J. Read, B. S. Sathyaprakash, D. Shoemaker, J. Smith, C. Torrie, S. Vitale, R. Weiss, C. Wipf, and M. Zucker (2020) "Cosmic Explorer: The U.S. Contribution to Gravitational-Wave Astronomy beyond LIGO," *Astro2020 APC White Papers*, **51**(7), `1907.04833`.

[41] Harry, I. W., B. Allen, and B. S. Sathyaprakash (2009) "A Stochastic template placement algorithm for gravitational wave data analysis," *Phys. Rev. D*, **80**, p. 104014, `0908.2090`.

[42] Privitera, S., S. R. P. Mohapatra, P. Ajith, K. Cannon, N. Fotopoulos, M. A. Frei, C. Hanna, A. J. Weinstein, and J. T. Whelan (2014) "Improving the sensitivity of a search for coalescing binary black holes with nonprecessing spins in gravitational wave data," *Phys. Rev. D*, **89**(2), p. 024003, `1310.5633`.

[43] Abbott, R. et al. (2023) "GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo during the Second Part of the Third Observing Run," *Phys. Rev. X*, **13**(4), p. 041039, `2111.03606`.

[44] Hanna, C. et al. (2023) "Binary tree approach to template placement for searches for gravitational waves from compact binary mergers," *Phys. Rev. D*, **108**(4), p. 042003, `2209.11298`.

[45] Sachdev, S. et al. (2019) "The GstLAL Search Analysis Methods for Compact Binary Mergers in Advanced LIGO's Second and Advanced Virgo's First Observing Runs," `1901.08580`.

[46] Cannon, K. et al. (2012) "Toward Early-Warning Detection of Gravitational Waves from Compact Binary Coalescence," *Astrophys. J.*, **748**, p. 136, `1107.2665`.

[47] Cannon, K., A. Chapman, C. Hanna, D. Keppel, A. C. Searle, and A. J. Weinstein (2010) "Singular value decomposition applied to compact binary coalescence gravitational-wave signals," *Phys. Rev. D*, **82**, p. 044025, `1005.0012`.

[48] Tsukada, L., K. Cannon, C. Hanna, D. Keppel, D. Meacher, and C. Messick (2018) "Application of a Zero-latency Whitening Filter to Compact Binary Coalescence Gravitational-wave Searches," *Phys. Rev. D*, **97**(10), p. 103009, `1708.04125`.

[49] Tsukada, L. et al. (2023) "Improved ranking statistics of the GstLAL inspiral search for compact binary coalescences," *Phys. Rev. D*, **108**(4), p. 043004, `2305.06286`.

[50] Cannon, K., C. Hanna, and J. Peoples (2015) "Likelihood-Ratio Ranking Statistic for Compact Binary Coalescence Candidates with Rate Estimation," `1504.04632`.

[51] Abbott, R. et al. (2021) "GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run," *Phys. Rev. X*, **11**, p. 021053, 2010.14527.

[52] ——— (2024) "GWTC-2.1: Deep extended catalog of compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run," *Phys. Rev. D*, **109**(2), p. 022001, 2108.01045.

[53] Biscans, S., S. Gras, C. D. Blair, J. Driggers, M. Evans, P. Fritschel, T. Hardwick, and G. Mansell (2019) "Suppressing parametric instabilities in LIGO using low-noise acoustic mode dampers," *Phys. Rev. D*, **100**(12), p. 122003, 1909.07805.

[54] Klimenko, S., I. Yakushin, A. Mercer, and G. Mitselmakher (2008) "Coherent method for detection of gravitational wave bursts," *Class. Quant. Grav.*, **25**, p. 114029, 0802.3232.

[55] Adams, T., D. Buskulic, V. Germain, G. M. Guidi, F. Marion, M. Montani, B. Mours, F. Piergiovanni, and G. Wang (2016) "Low-latency analysis pipeline for compact binary coalescences in the advanced gravitational wave detector era," *Class. Quant. Grav.*, **33**(17), p. 175012, 1512.02864.

[56] Dal Canton, T., A. H. Nitz, B. Gadre, G. S. Cabourn Davies, V. Villa-Ortega, T. Dent, I. Harry, and L. Xiao (2021) "Real-time Search for Compact Binary Mergers in Advanced LIGO and Virgo's Third Observing Run Using PyCBC Live," *Astrophys. J.*, **923**(2), p. 254, 2008.07494.

[57] Chu, Q. et al. (2022) "SPIIR online coherent pipeline to search for gravitational waves from compact binary coalescences," *Phys. Rev. D*, **105**(2), p. 024023, 2011.06787.

[58] Godwin, P. et al. (2020) "Incorporation of Statistical Data Quality Information into the GstLAL Search Analysis," 2010.15282.

[59] Abbott, R. et al. (2021) "Observation of Gravitational Waves from Two Neutron Star–Black Hole Coalescences," *Astrophys. J. Lett.*, **915**(1), p. L5, 2106.15163.

[60] Farr, W. M., N. Sravan, A. Cantrell, L. Kreidberg, C. D. Bailyn, I. Mandel, and V. Kalogera (2011) "The Mass Distribution of Stellar-Mass Black Holes," *Astrophys. J.*, **741**, p. 103, 1011.1459.

[61] Shao, D.-S., S.-P. Tang, J.-L. Jiang, and Y.-Z. Fan (2020) "Maximum mass cutoff in the neutron star mass distribution and the prospect of forming supramassive objects in the double neutron star mergers," *Phys. Rev. D*, **102**(6), p. 063006, 2009.04275.

[62] Abbott, R. et al. (2020) "GW190521: A Binary Black Hole Merger with a Total Mass of $150M_{\odot}$," *Phys. Rev. Lett.*, **125**(10), p. 101102, 2009.01075.

[63] Fishbach, M., R. Essick, and D. E. Holz (2020) "Does Matter Matter? Using the mass distribution to distinguish neutron stars and black holes," *Astrophys. J. Lett.*, **899**, p. L8, `2006.13178`.

[64] Mandel, I., W. M. Farr, A. Colonna, S. Stevenson, P. Tiňo, and J. Veitch (2017) "Model-independent inference on compact-binary observations," *Mon. Not. Roy. Astron. Soc.*, **465**(3), pp. 3254–3260, `1608.08223`.

[65] Chaudhary, S. S. et al. (2023) "Low-latency gravitational wave alert products and their performance in anticipation of the fourth LIGO-Virgo-KAGRA observing run," `2308.04545`.

[66] LIGO Scientific Collaboration and Virgo Collaboration (2023) *GCN*, **35000**.
URL `https://gcn.nasa.gov/circulars/35000.txt`

[67] ——— (2023) *GCN*, **34911**.
URL `https://gcn.nasa.gov/circulars/34911.txt`

[68] ——— (2023) *GCN*, **34728**.
URL `https://gcn.nasa.gov/circulars/34728.txt`

[69] ——— (2023) *GCN*, **34599**.
URL `https://gcn.nasa.gov/circulars/34599.txt`

[70] ——— (2023) *GCN*, **34378**.
URL `https://gcn.nasa.gov/circulars/34378.txt`

[71] ——— (2023) *GCN*, **34367**.
URL `https://gcn.nasa.gov/circulars/34367.txt`

[72] ——— (2023) *GCN*, **34600**.
URL `https://gcn.nasa.gov/circulars/34600.txt`

[73] ——— (2023) *GCN*, **34206**.
URL `https://gcn.nasa.gov/circulars/34206.txt`

[74] ——— (2023) *GCN*, **34172**.
URL `https://gcn.nasa.gov/circulars/34172.txt`

[75] ——— (2023) *GCN*, **34065**.
URL `https://gcn.nasa.gov/circulars/34065.txt`

[76] ——— (2023) *GCN*, **33871**.
URL `https://gcn.nasa.gov/circulars/33871.txt`

[77] ——— (2024) *GCN*, **35502**.
URL `https://gcn.nasa.gov/circulars/35502.txt`

[78] ——— (2024) *GCN*, **35493**.
URL https://gcn.nasa.gov/circulars/35493.txt

[79] ——— (2024) *GCN*, **35480**.
URL https://gcn.nasa.gov/circulars/35480.txt

[80] ——— (2023) *GCN*, **35445**.
URL https://gcn.nasa.gov/circulars/35445.txt

[81] ——— (2023) *GCN*, **35428**.
URL https://gcn.nasa.gov/circulars/35428.txt

[82] ——— (2024) *GCN*, **35423**.
URL https://gcn.nasa.gov/circulars/35423.txt

[83] ——— (2023) *GCN*, **35420**.
URL https://gcn.nasa.gov/circulars/35420.txt

[84] ——— (2023) *GCN*, **35330**.
URL https://gcn.nasa.gov/circulars/35330.txt

[85] ——— (2023) *GCN*, **35298**.
URL https://gcn.nasa.gov/circulars/35298.txt

[86] ——— (2023) *GCN*, **35297**.
URL https://gcn.nasa.gov/circulars/35297.txt

[87] ——— (2023) *GCN*, **35211**.
URL https://gcn.nasa.gov/circulars/35211.txt

[88] ——— (2023) *GCN*, **35202**.
URL https://gcn.nasa.gov/circulars/35202.txt

[89] ——— (2023) *GCN*, **35168**.
URL https://gcn.nasa.gov/circulars/35168.txt

[90] ——— (2023) *GCN*, **35120**.
URL https://gcn.nasa.gov/circulars/35120.txt

[91] ——— (2023) *GCN*, **35096**.
URL https://gcn.nasa.gov/circulars/35096.txt

[92] ——— (2023) *GCN*, **35094**.
URL https://gcn.nasa.gov/circulars/35094.txt

[93] ——— (2023) *GCN*, **35090**.
URL https://gcn.nasa.gov/circulars/35090.txt

[94] ——— (2023) *GCN*, **35023**.
URL `https://gcn.nasa.gov/circulars/35023.txt`

[95] ——— (2023) *GCN*, **35016**.
URL `https://gcn.nasa.gov/circulars/35016.txt`

[96] ——— (2023) *GCN*, **35010**.
URL `https://gcn.nasa.gov/circulars/35010.txt`

[97] ——— (2023) *GCN*, **34975**.
URL `https://gcn.nasa.gov/circulars/34975.txt`

[98] ——— (2023) *GCN*, **34967**.
URL `https://gcn.nasa.gov/circulars/34967.txt`

[99] ——— (2023) *GCN*, **34942**.
URL `https://gcn.nasa.gov/circulars/34942.txt`

[100] ——— (2023) *GCN*, **34927**.
URL `https://gcn.nasa.gov/circulars/34927.txt`

[101] ——— (2023) *GCN*, **34904**.
URL `https://gcn.nasa.gov/circulars/34904.txt`

[102] ——— (2023) *GCN*, **34895**.
URL `https://gcn.nasa.gov/circulars/34895.txt`

[103] ——— (2023) *GCN*, **34857**.
URL `https://gcn.nasa.gov/circulars/34857.txt`

[104] ——— (2023) *GCN*, **34852**.
URL `https://gcn.nasa.gov/circulars/34852.txt`

[105] ——— (2023) *GCN*, **34816**.
URL `https://gcn.nasa.gov/circulars/34816.txt`

[106] ——— (2023) *GCN*, **34807**.
URL `https://gcn.nasa.gov/circulars/34807.txt`

[107] ——— (2023) *GCN*, **34801**.
URL `https://gcn.nasa.gov/circulars/34801.txt`

[108] ——— (2023) *GCN*, **34799**.
URL `https://gcn.nasa.gov/circulars/34799.txt`

[109] ——— (2023) *GCN*, **34785**.
URL `https://gcn.nasa.gov/circulars/34785.txt`

[110] ——— (2023) *GCN*, **34783**.
URL https://gcn.nasa.gov/circulars/34783.txt

[111] ——— (2023) *GCN*, **34781**.
URL https://gcn.nasa.gov/circulars/34781.txt

[112] ——— (2023) *GCN*, **34775**.
URL https://gcn.nasa.gov/circulars/34775.txt

[113] ——— (2023) *GCN*, **34773**.
URL https://gcn.nasa.gov/circulars/34773.txt

[114] ——— (2024) *GCN*, **34760**.
URL https://gcn.nasa.gov/circulars/34760.txt

[115] ——— (2023) *GCN*, **34756**.
URL https://gcn.nasa.gov/circulars/34756.txt

[116] ——— (2023) *GCN*, **34757**.
URL https://gcn.nasa.gov/circulars/34757.txt

[117] ——— (2023) *GCN*, **34741**.
URL https://gcn.nasa.gov/circulars/34741.txt

[118] ——— (2023) *GCN*, **34739**.
URL https://gcn.nasa.gov/circulars/34739.txt

[119] ——— (2023) *GCN*, **34692**.
URL https://gcn.nasa.gov/circulars/34692.txt

[120] ——— (2023) *GCN*, **34666**.
URL https://gcn.nasa.gov/circulars/34666.txt

[121] ——— (2023) *GCN*, **34612**.
URL https://gcn.nasa.gov/circulars/34612.txt

[122] ——— (2023) *GCN*, **34605**.
URL https://gcn.nasa.gov/circulars/34605.txt

[123] ——— (2023) *GCN*, **34546**.
URL https://gcn.nasa.gov/circulars/34546.txt

[124] ——— (2024) *GCN*, **34534**.
URL https://gcn.nasa.gov/circulars/34534.txt

[125] ——— (2023) *GCN*, **34520**.
URL https://gcn.nasa.gov/circulars/34520.txt

[126] ——— (2023) *GCN*, **34504**.
URL `https://gcn.nasa.gov/circulars/34504.txt`

[127] ——— (2023) *GCN*, **34494**.
URL `https://gcn.nasa.gov/circulars/34494.txt`

[128] ——— (2023) *GCN*, **34429**.
URL `https://gcn.nasa.gov/circulars/34429.txt`

[129] ——— (2023) *GCN*, **34411**.
URL `https://gcn.nasa.gov/circulars/34411.txt`

[130] ——— (2023) *GCN*, **34380**.
URL `https://gcn.nasa.gov/circulars/34380.txt`

[131] ——— (2023) *GCN*, **34360**.
URL `https://gcn.nasa.gov/circulars/34360.txt`

[132] ——— (2023) *GCN*, **34352**.
URL `https://gcn.nasa.gov/circulars/34352.txt`

[133] ——— (2023) *GCN*, **34337**.
URL `https://gcn.nasa.gov/circulars/34337.txt`

[134] ——— (2023) *GCN*, **34314**.
URL `https://gcn.nasa.gov/circulars/34314.txt`

[135] ——— (2023) *GCN*, **34303**.
URL `https://gcn.nasa.gov/circulars/34303.txt`

[136] ——— (2023) *GCN*, **34293**.
URL `https://gcn.nasa.gov/circulars/34293.txt`

[137] ——— (2023) *GCN*, **34264**.
URL `https://gcn.nasa.gov/circulars/34264.txt`

[138] ——— (2023) *GCN*, **34235**.
URL `https://gcn.nasa.gov/circulars/34235.txt`

[139] ——— (2023) *GCN*, **34175**.
URL `https://gcn.nasa.gov/circulars/34175.txt`

[140] ——— (2023) *GCN*, **34173**.
URL `https://gcn.nasa.gov/circulars/34173.txt`

[141] ——— (2023) *GCN*, **34195**.
URL `https://gcn.nasa.gov/circulars/34195.txt`

[142] ——— (2023) *GCN*, **34194**.
URL `https://gcn.nasa.gov/circulars/34194.txt`

[143] ——— (2023) *GCN*, **34161**.
URL `https://gcn.nasa.gov/circulars/34161.txt`

[144] ——— (2023) *GCN*, **34147**.
URL `https://gcn.nasa.gov/circulars/34147.txt`

[145] ——— (2023) *GCN*, **34140**.
URL `https://gcn.nasa.gov/circulars/34140.txt`

[146] ——— (2023) *GCN*, **34136**.
URL `https://gcn.nasa.gov/circulars/34136.txt`

[147] ——— (2023) *GCN*, **34127**.
URL `https://gcn.nasa.gov/circulars/34127.txt`

[148] ——— (2023) *GCN*, **34124**.
URL `https://gcn.nasa.gov/circulars/34124.txt`

[149] ——— (2023) *GCN*, **34113**.
URL `https://gcn.nasa.gov/circulars/34113.txt`

[150] ——— (2023) *GCN*, **34086**.
URL `https://gcn.nasa.gov/circulars/34086.txt`

[151] ——— (2024) *GCN*, **34075**.
URL `https://gcn.nasa.gov/circulars/34075.txt`

[152] ——— (2023) *GCN*, **33944**.
URL `https://gcn.nasa.gov/circulars/33944.txt`

[153] ——— (2023) *GCN*, **33938**.
URL `https://gcn.nasa.gov/circulars/33938.txt`

[154] ——— (2023) *GCN*, **33922**.
URL `https://gcn.nasa.gov/circulars/33922.txt`

[155] ——— (2023) *GCN*, **33914**.
URL `https://gcn.nasa.gov/circulars/33914.txt`

[156] ——— (2023) *GCN*, **33903**.
URL `https://gcn.nasa.gov/circulars/33903.txt`

[157] ——— (2023) *GCN*, **33889**.
URL `https://gcn.nasa.gov/circulars/33889.txt`

[158] Biscoveanu, S., P. Landry, and S. Vitale (2022) "Population properties and multimessenger prospects of neutron star–black hole mergers following GWTC-3," *Mon. Not. Roy. Astron. Soc.*, **518**(4), pp. 5298–5312, 2207.01568.

[159] Abbott, R. et al. (2020) "GW190814: Gravitational Waves from the Coalescence of a 23 Solar Mass Black Hole with a 2.6 Solar Mass Compact Object," *Astrophys. J. Lett.*, **896**(2), p. L44, 2006.12611.

[160] Pratten, G. et al. (2021) "Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes," *Phys. Rev. D*, **103**(10), p. 104056, 2004.06503.

[161] Husa, S., S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. Jiménez Forteza, and A. Bohé (2016) "Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal," *Phys. Rev. D*, **93**(4), p. 044006, 1508.07250.

[162] Davis, D., M. Trevor, S. Mozzon, and L. K. Nuttall (2022) "Incorporating information from LIGO data quality streams into the PyCBC search for gravitational waves," *Phys. Rev. D*, **106**(10), p. 102006, 2204.03091.

[163] Scientific), F. A. V. (2015) "Advanced Virgo: a second-generation interferometric gravitational wave detector," *Classical and Quantum Gravity*, 1408.3978.

[164] Nuttall, L. K. (2018) "Characterizing transient noise in the LIGO detectors," *Phil. Trans. Roy. Soc. Lond. A*, **376**(2120), p. 20170286, 1804.07592.

[165] Cabero, M. et al. (2019) "Blip glitches in Advanced LIGO data," *Class. Quant. Grav.*, **36**(15), p. 15, 1901.05093.

[166] Abbott, B. P. et al. (2020) "A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals," *Class. Quant. Grav.*, **37**(5), p. 055002, 1908.11170.

[167] ——— (2016) "Characterization of transient noise in Advanced LIGO relevant to gravitational wave signal GW150914," *Class. Quant. Grav.*, **33**(13), p. 134001, 1602.03844.

[168] Canton, T. D., S. Bhagwat, S. V. Dhurandhar, and A. Lundgren (2014) "Effect of sine-Gaussian glitches on searches for binary coalescence," *Class. Quant. Grav.*, **31**, p. 015016, 1304.0008.

[169] Abbott, B. P. et al. (2018) "Effects of data quality vetoes on a search for compact binary coalescences in Advanced LIGO's first observing run," *Class. Quant. Grav.*, **35**(6), p. 065010, 1710.02185.

[170] Powell, J. (2018) "Parameter Estimation and Model Selection of Gravitational Wave Signals Contaminated by Transient Detector Noise Glitches," *Class. Quant. Grav.*, **35**(15), p. 155017, 1803.11346.

[171] The LIGO Scientific Collaboration, T. V. C. (2017) "GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral," *Physical Review Letters*, 1710.05832.

[172] Pankow, C. et al. (2018) "Mitigation of the instrumental noise transient in gravitational-wave data surrounding GW170817," *Phys. Rev. D*, **98**(8), p. 084016, 1808.03619.

[173] Matichard, F. et al. (2015) "Seismic isolation of Advanced LIGO: Review of strategy, instrumentation and performance," *Class. Quant. Grav.*, **32**(18), p. 185003, 1502.06300.

[174] Graef Rollins, J. (2016) "Distributed state machine supervision for long-baseline gravitational-wave detectors," *Rev. Sci. Instrum.*, **87**(9), p. 094502.

[175] Mueller, C. L. et al. (2016) "The Advanced LIGO Input Optics," *Rev. Sci. Instrum.*, **87**(1), p. 014502, 1601.05442.

[176] Effler, A., R. M. S. Schofield, V. V. Frolov, G. González, K. Kawabe, J. R. Smith, J. Birch, and R. McCarthy (2015) "Environmental Influences on the LIGO Gravitational Wave Detectors during the 6th Science Run," *Class. Quant. Grav.*, **32**(3), p. 035017, 1409.5160.

[177] Soni, S. et al. (2021) "Discovering features in gravitational-wave data through detector characterization, citizen science and machine learning," *Class. Quant. Grav.*, **38**(19), p. 195016, 2103.12104.

[178] Smith, J. R., T. Abbott, E. Hirose, N. Leroy, D. Macleod, J. McIver, P. Saulson, and P. Shawhan (2011) "A Hierarchical method for vetoing noise transients in gravitational-wave detectors," *Class. Quant. Grav.*, **28**, p. 235005, 1107.2948.

[179] Lattimer, J. M. and M. Prakash (2016) "The Equation of State of Hot, Dense Matter and Neutron Stars," *Phys. Rept.*, **621**, pp. 127–164, 1512.07820.

[180] Özel, F. and P. Freire (2016) "Masses, Radii, and the Equation of State of Neutron Stars," *Ann. Rev. Astron. Astrophys.*, **54**, pp. 401–440, 1603.02698.

[181] Baym, G., T. Hatsuda, T. Kojo, P. D. Powell, Y. Song, and T. Takatsuka (2018) "From hadrons to quarks in neutron stars: a review," *Rep. Prog. Phys.*, **81**(5), p. 056902.
URL http://dx.doi.org/10.1088/1361-6633/aaae14

[182] Raaijmakers, G. et al. (2020) "Constraining the dense matter equation of state with joint analysis of NICER and LIGO/Virgo measurements," *Astrophys. J. Lett.*, **893**(1), p. L21, 1912.11031.

[183] Raaijmakers, G., S. K. Greif, K. Hebeler, T. Hinderer, S. Nissanke, A. Schwenk, T. E. Riley, A. L. Watts, J. M. Lattimer, and W. C. G. Ho (2021) "Constraints on the Dense Matter Equation of State and Neutron Star Properties from NICER's Mass–Radius Estimate of PSR J0740+6620 and Multimessenger Observations," *Astrophys. J. Lett.*, **918**, p. L29.
URL http://dx.doi.org/10.3847/2041-8213/ac089a

[184] Ayriyan, A., D. Blaschke, A. G. Grunfeld, D. Alvarez-Castillo, H. Grigorian, and V. Abgaryan (2021) "Bayesian analysis of multimessenger M-R data with interpolated hybrid EoS," *Eur. Phys. J. A*, **57**(11), p. 318, 2102.13485.

[185] Essick, R., I. Tews, P. Landry, S. Reddy, and D. E. Holz (2020) "Direct Astrophysical Tests of Chiral Effective Field Theory at Supranuclear Densities," *Phys. Rev. C*, **102**(5), p. 055803, 2004.07744.

[186] Hu, J., S. Bao, Y. Zhang, K. Nakazato, K. Sumiyoshi, and H. Shen (2020) "Effects of symmetry energy on the radius and tidal deformability of neutron stars in the relativistic mean-field model," *PTEP*, **2020**(4), p. 043D01, 2002.00562.

[187] Raithel, C. A. (2019) "Constraints on the Neutron Star Equation of State from GW170817," *Eur. Phys. J. A*, **55**(5), p. 80, 1904.10002.

[188] Miller, M. C. et al. (2021) "The Radius of PSR J0740+6620 from NICER and XMM-Newton Data," *Astrophys. J. Lett.*, **918**(2), p. L28, 2105.06979.

[189] Han, S., M. A. A. Mamun, S. Lalit, C. Constantinou, and M. Prakash (2019) "Treating quarks within neutron stars," *Phys. Rev. D*, **100**(10), p. 103022, 1906.04095.

[190] Li, A., Z. Miao, S. Han, and B. Zhang (2021) "Constraints on the maximum mass of neutron stars with a quark core from GW170817 and NICER PSR J0030+0451 data," *Astrophys. J.*, **913**(1), p. 27, 2103.15119.

[191] Christian, J.-E., A. Zacchi, and J. Schaffner-Bielich (2019) "Signals in the tidal deformability for phase transitions in compact stars with constraints from GW170817," *Phys. Rev. D*, **99**(2), p. 023009, 1809.03333.

[192] Montana, G., L. Tolos, M. Hanauske, and L. Rezzolla (2019) "Constraining twin stars with GW170817," *Phys. Rev. D*, **99**(10), p. 103009, 1811.10929.

[193] Li, C.-M., Y. Yan, J.-J. Geng, Y.-F. Huang, and H.-S. Zong (2018) "Constraints on the hybrid equation of state with a crossover hadron-quark phase transition in the light of GW170817," *Phys. Rev. D*, **98**(8), p. 083013, 1808.02601.

[194] Blaschke, D. and M. Cierniak (2021) "Studying the onset of deconfinement with multi-messenger astronomy of neutron stars," *Astron. Nachr.*, **342**(1-2), pp. 227–233, 2012.15785.

[195] Postnov, K. A. and L. R. Yungelson (2014) "The Evolution of Compact Binary Star Systems," *Living Rev. Relativ.*, **17**(1), p. 3.
URL http://dx.doi.org/10.12942/lrr-2014-3

[196] Steiner, A. W. and S. Gandolfi (2012) "Connecting Neutron Star Observations to Three-Body Forces in Neutron Matter and to the Nuclear Symmetry Energy," *Phys. Rev. Lett.*, **108**, p. 081102, 1110.4142.

[197] Catuneanu, A., C. O. Heinke, G. R. Sivakoff, W. C. G. Ho, and M. Servillat (2013) "Mass/Radius Constraints on the Quiescent Neutron Star in M13 Using Hydrogen and Helium Atmospheres," *Astrophys. J.*, **764**, p. 145, 1301.3768.

[198] Guver, T., P. Wroblewski, L. Camarota, and F. Ozel (2010) "The Mass and Radius of the Neutron Star in 4U 1820-30," *Astrophys. J.*, **719**, p. 1807, 1002.3825.

[199] Guver, T., F. Ozel, A. Cabrera-Lavers, and P. Wroblewski (2010) "The Distance, Mass, and Radius of the Neutron Star in 4U 1608-52," *Astrophys. J.*, **712**, pp. 964–973, 0811.3979.

[200] Lattimer, J. M. and A. W. Steiner (2014) "Neutron Star Masses and Radii from Quiescent Low-Mass X-ray Binaries," *Astrophys. J.*, **784**, p. 123, 1305.3242.

[201] Poutanen, J., J. Nättilä, J. J. E. Kajava, O.-M. Latvala, D. Galloway, E. Kuulkers, and V. Suleimanov (2014) "The effect of accretion on the measurement of neutron star mass and radius in the low-mass X-ray binary 4U 1608−52," *Mon. Not. Roy. Astron. Soc.*, **442**(4), pp. 3777–3790, 1405.2663.

[202] Gendreau, K. C., Z. Arzoumanian, P. W. Adkins, C. L. Albert, J. F. Anders, A. T. Aylward, C. L. Baker, E. R. Balsamo, W. A. Bamford, S. S. Benegalrao, D. L. Berry, S. Bhalwani, J. K. Black, C. Blaurock, G. M. Bronke, G. L. Brown, J. G. Budinoff, J. D. Cantwell, T. Cazeau, P. T. Chen, T. G. Clement, A. T. Colangelo, J. S. Coleman, J. D. Coopersmith, W. E. Dehaven, J. P. Doty, M. D. Egan, T. Enoto, T. W. Fan, D. M. Ferro, R. Foster, N. M. Galassi, L. D. Gallo, C. M. Green, D. Grosh, K. Q. Ha, M. A. Hasouneh, K. B. Heefner, P. Hestnes, L. J. Hoge, T. M. Jacobs, J. L. Jørgensen, M. A. Kaiser, J. W. Kellogg, S. J. Kenyon, R. G. Koenecke, R. P. Kozon, B. LaMarr, M. D. Lambertson, A. M. Larson, S. Lentine, J. H. Lewis, M. G. Lilly, K. A. Liu, A. Malonis, S. S. Manthripragada, C. B. Markwardt, B. D. Matonak, I. E. Mcginnis, R. L. Miller, A. L. Mitchell, J. W. Mitchell, J. S. Mohammed, C. A. Monroe, K. M. Montt de Garcia, P. D. Mulé, L. T. Nagao, S. N. Ngo, E. D. Norris, D. A. Norwood, J. Novotka, T. Okajima, L. G. Olsen, C. O. Onyeachu, H. Y. Orosco, J. R. Peterson, K. N. Pevear, K. K. Pham, S. E. Pollard, J. S. Pope, D. F. Powers, C. E. Powers, S. R. Price, G. Y. Prigozhin, J. B. Ramirez, W. J. Reid, R. A. Remillard, E. M. Rogstad, G. P. Rosecrans, J. N. Rowe, J. A. Sager, C. A. Sanders, B. Savadkin, M. R. Saylor, A. F. Schaeffer, N. S. Schweiss, S. R. Semper, P. J. Serlemitsos, L. V. Shackelford, Y. Soong, J. Struebel, M. L. Vezie, J. S. Villasenor, L. B. Winternitz, G. I. Wofford, M. R. Wright, M. Y. Yang, and W. H. Yu (2016) "The Neutron star Interior Composition

Explorer (NICER): design and development," in *Space Telescopes and Instrumentation 2016: Ultraviolet to Gamma Ray* (J.-W. A. den Herder, T. Takahashi, and M. Bautz, eds.), vol. 9905, SPIE, pp. 420–435.
URL `https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9905/99051H/The--Neutron-star-Interior-Composition-Explorer-NICER--design/10.1117/12.2231304.short`

[203] Miller, M. C. et al. (2019) "PSR J0030+0451 Mass and Radius from *NICER* Data and Implications for the Properties of Neutron Star Matter," *Astrophys. J. Lett.*, **887**(1), p. L24, `1912.05705`.

[204] Raaijmakers, G. et al. (2019) "A *NICER* view of PSR J0030+0451: Implications for the dense matter equation of state," *Astrophys. J. Lett.*, **887**(1), p. L22, `1912.05703`.

[205] Bogdanov, S. et al. (2019) "Constraining the Neutron Star Mass–Radius Relation and Dense Matter Equation of State with *NICER*. I. The Millisecond Pulsar X-Ray Data Set," *Astrophys. J. Lett.*, **887**(1), p. L25, `1912.05706`.

[206] ——— (2019) "Constraining the Neutron Star Mass–Radius Relation and Dense Matter Equation of State with *NICER*. II. Emission from Hot Spots on a Rapidly Rotating Neutron Star," *Astrophys. J. Lett.*, **887**(1), p. L26, `1912.05707`.

[207] Watts, A. L. (2019) "Constraining the neutron star equation of state using Pulse Profile Modeling," *AIP Conf. Proc.*, **2127**(1), p. 020008, `1904.07012`.

[208] Aasi, J. et al. (2015) "Advanced LIGO," *Class. Quant. Grav.*, **32**, p. 074001, `1411.4547`.

[209] Acernese, F. et al. (2015) "Advanced Virgo: a second-generation interferometric gravitational wave detector," *Class. Quant. Grav.*, **32**(2), p. 024001, `1408.3978`.

[210] Abbott, B. P. et al. (2019) "Properties of the binary neutron star merger GW170817," *Phys. Rev. X*, **9**(1), p. 011001, `1805.11579`.

[211] ——— (2019) "GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs," *Phys. Rev. X*, **9**(3), p. 031040, `1811.12907`.

[212] ——— (2020) "GW190425: Observation of a Compact Binary Coalescence with Total Mass $\sim 3.4 M_{\odot}$," *Astrophys. J. Lett.*, **892**(1), p. L3, `2001.01761`.

[213] Lai, D., F. A. Rasio, and S. L. Shapiro (1994) "Hydrodynamic instability and coalescence of binary neutron stars," *Astrophys. J.*, **420**, pp. 811–829, `astro-ph/9304027`.

[214] Cutler, C. and E. E. Flanagan (1994) "Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral wave form?" *Phys. Rev. D*, **49**, pp. 2658–2697, `gr-qc/9402014`.

[215] Kokkotas, K. D. and B. G. Schmidt (1999) "Quasinormal modes of stars and black holes," *Living Rev. Rel.*, **2**, p. 2, gr-qc/9909058.

[216] Flanagan, É. É. and T. Hinderer (2008) "Constraining neutron-star tidal Love numbers with gravitational-wave detectors," *Phys. Rev. D*, **77**(2), p. 021502.
URL https://link.aps.org/doi/10.1103/PhysRevD.77.021502

[217] Hinderer, T., B. D. Lackey, R. N. Lang, and J. S. Read (2010) "Tidal deformability of neutron stars with realistic equations of state and their gravitational wave signatures in binary inspiral," *Phys. Rev. D*, **81**(12), p. 123016.
URL https://link.aps.org/doi/10.1103/PhysRevD.81.123016

[218] Hinderer, T. (2008) "Tidal love numbers of neutron stars," *Astrophys. J.*, **677**(2), pp. 1216–1220.
URL https://iopscience.iop.org/article/10.1086/533487

[219] Vines, J., É. É. Flanagan, and T. Hinderer (2011) "Post-1-Newtonian tidal effects in the gravitational waveform from binary inspirals," *Phys. Rev. D*, **83**(8), p. 084051.
URL https://link.aps.org/doi/10.1103/PhysRevD.83.084051

[220] Harry, I. and A. Lundgren (2021) "Failure of the Fisher matrix when including tidal terms: Considering construction of template banks of tidally deformed binary neutron stars," *Phys. Rev. D*, **104**(4), p. 043008, 2101.01091.

[221] Favata, M. (2014) "Systematic parameter errors in inspiraling neutron star binaries," *Phys. Rev. Lett.*, **112**(10), p. 101101.
URL http://dx.doi.org/10.1103/PhysRevLett.112.101101

[222] Hinderer, T. et al. (2016) "Effects of neutron-star dynamic tides on gravitational waveforms within the effective-one-body approach," *Phys. Rev. Lett.*, **116**(18), p. 181101, 1602.00599.

[223] Dietrich, T. and T. Hinderer (2017) "Comprehensive comparison of numerical relativity and effective-one-body results to inform improvements in waveform models for binary neutron star systems," *Phys. Rev. D*, **95**(12), p. 124006, 1702.02053.

[224] Dietrich, T., S. Bernuzzi, and W. Tichy (2017) "Closed-form tidal approximants for binary neutron star gravitational waveforms constructed from high-resolution numerical relativity simulations," *Phys. Rev. D*, **96**(12), p. 121501.
URL https://link.aps.org/doi/10.1103/PhysRevD.96.121501

[225] Dietrich, T., S. Khan, R. Dudi, S. J. Kapadia, P. Kumar, A. Nagar, F. Ohme, F. Pannarale, A. Samajdar, S. Bernuzzi, G. Carullo, W. Del Pozzo, M. Haney, C. Markakis, M. Puerrer, G. Riemenschneider, Y. E. Setyawati, K. W. Tsang, and C. Van Den Broeck (2018) "Matter imprints in waveform models for neutron star binaries: tidal and self-spin effects," *arXiv [gr-qc]*, 1804.02235.
URL http://arxiv.org/abs/1804.02235

[226] Dietrich, T., A. Samajdar, S. Khan, N. K. Johnson-McDaniel, R. Dudi, and W. Tichy (2019) "Improving the NRTidal model for binary neutron star systems," *Phys. Rev. D*, **100**(4), p. 044003.
URL https://link.aps.org/doi/10.1103/PhysRevD.100.044003

[227] Nagar, A. et al. (2018) "Time-domain effective-one-body gravitational waveforms for coalescing compact binaries with nonprecessing spins, tides and self-spin effects," *Phys. Rev. D*, **98**(10), p. 104052, 1806.01772.

[228] Henry, Q., G. Faye, and L. Blanchet (2020) "Hamiltonian for tidal interactions in compact binary systems to next-to-next-to-leading post-Newtonian order," *Phys. Rev. D*, **102**(12), p. 124074, 2009.12332.

[229] ——— (2020) "Tidal effects in the gravitational-wave phase evolution of compact binary systems to next-to-next-to-leading post-Newtonian order," *Phys. Rev. D*, **102**(4), p. 044033, 2005.13367.

[230] Wade, L., J. D. E. Creighton, E. Ochsner, B. D. Lackey, B. F. Farr, T. B. Littenberg, and V. Raymond (2014) "Systematic and statistical errors in a bayesian approach to the estimation of the neutron-star equation of state using advanced gravitational wave detectors," *Phys. Rev. D*, **89**(10), p. 103012, 1402.5156.

[231] Riley, T. E., G. Raaijmakers, and A. L. Watts (2018) "On parametrized cold dense matter equation-of-state inference," *Mon. Not. Roy. Astron. Soc.*, **478**(1), pp. 1093–1131, 1804.09085.

[232] De, S., D. Finstad, J. M. Lattimer, D. A. Brown, E. Berger, and C. M. Biwer (2018) "Tidal Deformabilities and Radii of Neutron Stars from the Observation of GW170817," *Phys. Rev. Lett.*, **121**(9), p. 091102, [Erratum: Phys.Rev.Lett. 121, 259902 (2018)], 1804.08583.

[233] The LIGO Scientific Collaboration, T. V. C. (2020) "GW190425: Observation of a Compact Binary Coalescence with Total Mass ~ 3.4 Solar Mass," *Astrophysical Journal Letters*, **892**(1), 2001.01761.

[234] Capano, C. D., I. Tews, S. M. Brown, B. Margalit, S. De, S. Kumar, D. A. Brown, B. Krishnan, and S. Reddy (2020) "Stringent constraints on neutron-star radii from multimessenger observations and nuclear theory," *Nature Astron.*, **4**(6), pp. 625–632, 1908.10352.

[235] Biswas, B. (2021) "Impact of PREX-II and Combined Radio/NICER/XMM-Newton's Mass–radius Measurement of PSR J0740+6620 on the Dense-matter Equation of State," *ApJ*, **921**(1), p. 63.
URL https://iopscience.iop.org/article/10.3847/1538-4357/ac1c72/meta

[236] Breschi, M., A. Perego, S. Bernuzzi, W. Del Pozzo, V. Nedora, D. Radice, and D. Vescovi (2021) "AT2017gfo: Bayesian inference and model selection of multi-component kilonovae and constraints on the neutron star equation of state," *Mon. Not. Roy. Astron. Soc.*, **505**(2), pp. 1661–1677, 2101.01201.

[237] Vinciguerra, S. et al. (2023) "An updated mass-radius analysis of the 2017-2018 NICER data set of PSR J0030+0451," 2308.09469.

[238] Puecher, A., A. Samajdar, and T. Dietrich (2023) "Measuring tidal effects with the Einstein Telescope: A design study," 2304.05349.

[239] Pacilio, C., A. Maselli, M. Fasano, and P. Pani (2022) "Ranking Love Numbers for the Neutron Star Equation of State: The Need for Third-Generation Detectors," *Phys. Rev. Lett.*, **128**(10), p. 101101, 2104.10035.

[240] Ghosh, T., B. Biswas, and S. Bose (2022) "Simultaneous inference of neutron star equation of state and the Hubble constant with a population of merging neutron stars," *Phys. Rev. D*, **106**(12), p. 123529.
URL https://link.aps.org/doi/10.1103/PhysRevD.106.123529

[241] Chatterjee, D., A. Hegade K. R., G. Holder, D. E. Holz, S. Perkins, K. Yagi, and N. Yunes (2021) "Cosmology with Love: Measuring the Hubble constant using neutron star universal relations," *Phys. Rev. D*, **104**(8), p. 083528.
URL https://link.aps.org/doi/10.1103/PhysRevD.104.083528

[242] Beniamini, P. and T. Piran (2019) "The Gravitational waves merger time distribution of binary neutron star systems," *Mon. Not. Roy. Astron. Soc.*, **487**(4), pp. 4847–4854, 1903.11614.

[243] Greggio, L., P. Simonetti, and F. Matteucci (2020) "On the delay times of merging double neutron stars," *Mon. Not. Roy. Astron. Soc.*, **500**(2), pp. 1755–1771, 2009.13138.

[244] Vangioni, E., K. A. Olive, T. Prestegard, J. Silk, P. Petitjean, and V. Mandic (2015) "The Impact of Star Formation and Gamma-Ray Burst Rates at High Redshift on Cosmic Chemical Evolution and Reionization," *Mon. Not. Roy. Astron. Soc.*, **447**, p. 2575, 1409.2462.

[245] Hannam, M., P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer (2014) "Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms," *Phys. Rev. Lett.*, **113**(15), p. 151101, 1308.3271.

[246] Khan, S., S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé (2016) "Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era," *Phys. Rev. D*, **93**(4), p. 044007, 1508.07253.

[247] Godzieba, D. A., R. Gamba, D. Radice, and S. Bernuzzi (2021) "Updated universal relations for tidal deformabilities of neutron stars from phenomenological equations of state," *Phys. Rev. D*, **103**(6), p. 063036, 2012.12151.

[248] Antoniadis, J. et al. (2013) "A Massive Pulsar in a Compact Relativistic Binary," *Science*, **340**, p. 6131, 1304.6875.

[249] Fonseca, E. et al. (2021) "Refined Mass and Geometric Measurements of the High-mass PSR J0740+6620," *Astrophys. J. Lett.*, **915**(1), p. L12, 2104.00880.

[250] Romani, R. W., D. Kandel, A. V. Filippenko, T. G. Brink, and W. Zheng (2022) "PSR J0952−0607: The Fastest and Heaviest Known Galactic Neutron Star," *Astrophys. J. Lett.*, **934**(2), p. L18, 2207.05124.

[251] Oppenheimer, J. R. and G. M. Volkoff (1939) "On massive neutron cores," *Phys. Rev.*, **55**, pp. 374–381.

[252] Tolman, R. C. (1939) "Static solutions of Einstein's field equations for spheres of fluid," *Phys. Rev.*, **55**, pp. 364–373.

[253] Adhikari, R. X. et al. (2019) "Astrophysical science metrics for next-generation gravitational-wave detectors," *Class. Quant. Grav.*, **36**(24), p. 245010, 1905.02842.

[254] Borhanian, S. (2021) "GWBENCH: a novel Fisher information package for gravitational-wave benchmarking," *Class. Quant. Grav.*, **38**(17), p. 175014, 2010.15202.

[255] Arun, K. G., B. R. Iyer, B. S. Sathyaprakash, and P. A. Sundararajan (2005) "Parameter estimation of inspiralling compact binaries using 3.5 post-Newtonian gravitational wave phasing: The Non-spinning case," *Phys. Rev. D*, **71**, p. 084008, [Erratum: Phys.Rev.D 72, 069903 (2005)], gr-qc/0411146.

[256] Yagi, K. and N. Yunes (2017) "Approximate Universal Relations for Neutron Stars and Quark Stars," *Phys. Rept.*, **681**, pp. 1–72, 1608.02582.

[257] Chatziioannou, K., C.-J. Haster, and A. Zimmerman (2018) "Measuring the neutron star tidal deformability with equation-of-state-independent relations and gravitational waves," *Phys. Rev. D*, **97**(10), p. 104036, 1804.03221.

[258] Maselli, A., V. Cardoso, V. Ferrari, L. Gualtieri, and P. Pani (2013) "Equation-of-state-independent relations in neutron stars," *Phys. Rev. D*, **88**(2), p. 023007, 1304.2052.

[259] Pradhan, B. K., A. Vijaykumar, and D. Chatterjee (2023) "Impact of updated multipole Love numbers and f-Love universal relations in the context of binary neutron stars," *Phys. Rev. D*, **107**(2), p. 023010, 2210.09425.

[260] Kastaun, W. and F. Ohme (2019) "Finite tidal effects in GW170817: Observational evidence or model assumptions?" *Phys. Rev. D*, **100**(10), p. 103023, 1909.12718.

[261] Kashyap, R., A. Dhani, and B. Sathyaprakash (2022) “Systematic errors due to quasiuniversal relations in binary neutron stars and their correction for unbiased model selection,” *Phys. Rev. D*, **106**(12), p. 123001, 2209.02757.

[262] Rose, H., N. Kunert, T. Dietrich, P. T. H. Pang, R. Smith, C. Van Den Broeck, S. Gandolfi, and I. Tews (2023) “Revealing the strength of three-nucleon interactions with the Einstein Telescope,” 2303.11201.

[263] Sabatucci, A., O. Benhar, A. Maselli, and C. Pacilio (2022) “Sensitivity of neutron star observations to three-nucleon forces,” *Phys. Rev. D*, **106**(8), p. 083010, 2206.11286.

[264] Borhanian, S., A. Dhani, A. Gupta, K. G. Arun, and B. S. Sathyaprakash (2020) “Dark Sirens to Resolve the Hubble–Lemaître Tension,” *Astrophys. J. Lett.*, **905**(2), p. L28, 2007.02883.

[265] Gupta, I. et al. (2023) “Characterizing Gravitational Wave Detector Networks: From A$^\sharp$ to Cosmic Explorer,” 2307.10421.

[266] Gupta, I. (2023) “Using grey sirens to resolve the Hubble–Lemaître tension,” *Mon. Not. Roy. Astron. Soc.*, **524**(3), pp. 3537–3558, 2212.00163.

[267] Branchesi, M. et al. (2023) “Science with the Einstein Telescope: a comparison of different designs,” *JCAP*, **07**, p. 068, 2303.15923.

[268] Dhani, A., S. Borhanian, A. Gupta, and B. Sathyaprakash (2022) “Cosmography with bright and Love sirens,” 2212.13183.

[269] Muttoni, N., D. Laghi, N. Tamanini, S. Marsat, and D. Izquierdo-Villalba (2023) “Dark siren cosmology with binary black holes in the era of third-generation gravitational wave detectors,” *Phys. Rev. D*, **108**(4), p. 043543, 2303.10693.

[270] Aghanim, N. et al. (2020) “Planck 2018 results. VI. Cosmological parameters,” *Astron. Astrophys.*, **641**, p. A6, [Erratum: Astron.Astrophys. 652, C4 (2021)], 1807.06209.

[271] Riess, A. G. et al. (2022) “A Comprehensive Measurement of the Local Value of the Hubble Constant with 1 km s^{-1} Mpc^{-1} Uncertainty from the Hubble Space Telescope and the SH0ES Team,” *Astrophys. J. Lett.*, **934**(1), p. L7, 2112.04510.

[272] Ashton, G. et al. (2019) “BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy,” *Astrophys. J. Suppl.*, **241**(2), p. 27, 1811.02042.

[273] Romero-Shaw, I. M. et al. (2020) “Bayesian inference for compact binary coalescences with bilby: validation and application to the first LIGO–Virgo gravitational-wave transient catalogue,” *Mon. Not. Roy. Astron. Soc.*, **499**(3), pp. 3295–3319, 2006.00714.

[274] Abbott, B. P. and Others (2020) “Model comparison from LIGO-Virgo data on GW170817’s binary components and consequences for the merger remnant,” *Classical Quantum Gravity*, **37**(4), p. 045006, 1908.01012.
URL http://dx.doi.org/10.1088/1361-6382/ab5f7c

[275] Ghosh, S., X. Liu, J. Creighton, W. Kastaun, G. Pratten, and I. M. Hernandez (2021) "Rapid model comparison of equations of state from gravitational wave observation of binary neutron star coalescences," *Phys. Rev. D*, **104**(8), p. 083003, 2104.08681.

[276] Biswas, B. (2022) "Bayesian Model Selection of Neutron Star Equations of State Using Multi-messenger Observations," *Astrophys. J.*, **926**(1), p. 75, 2106.02644.

Vita

Rachael Huxford

Education

The Pennsylvania State University, Ph.D. Physics	2024
Towson University, B.Sc. Physics	2018

Selected Publications

Huxford, Rachael et. al., "The Accuracy of Neutron Star Radius Measurement with the Next Generation of Terrestrial Gravitational-Wave Observatories"

Huxford, Rachael et. al., "Performance of iDQ ahead of LIGO, Virgo, and KAGRA's fourth observing run" (In Preparation)

Ewing, Becca et. al. "Performance of the low-latency GstLAL inspiral search towards LIGO, Virgo, and KAGRA's fourth observing run"

The LIGO Scientific Collaboration et. al., "GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run"

The LIGO Scientific Collaboration et. al., "Observation of Gravitational Waves from the Coalescence of a 2.5–4.5$M_\odot$ Compact Object and a Neutron Star”

Selected Awards and Fellowships

George A. and Margaret M. Downsbrough Department Head's Chair in Physics	2020
Nellie H. and Oscar L. Roberts Scholarship Fund	2019
Paul Berg and Daniel J. Larson Distinguished Graduate Fellowship	2019
Edward I. Rubendall Physics Achievement Award for Outstanding Senior	2018