Data Analysis of Gravitational Waves: Matched Filtering Method

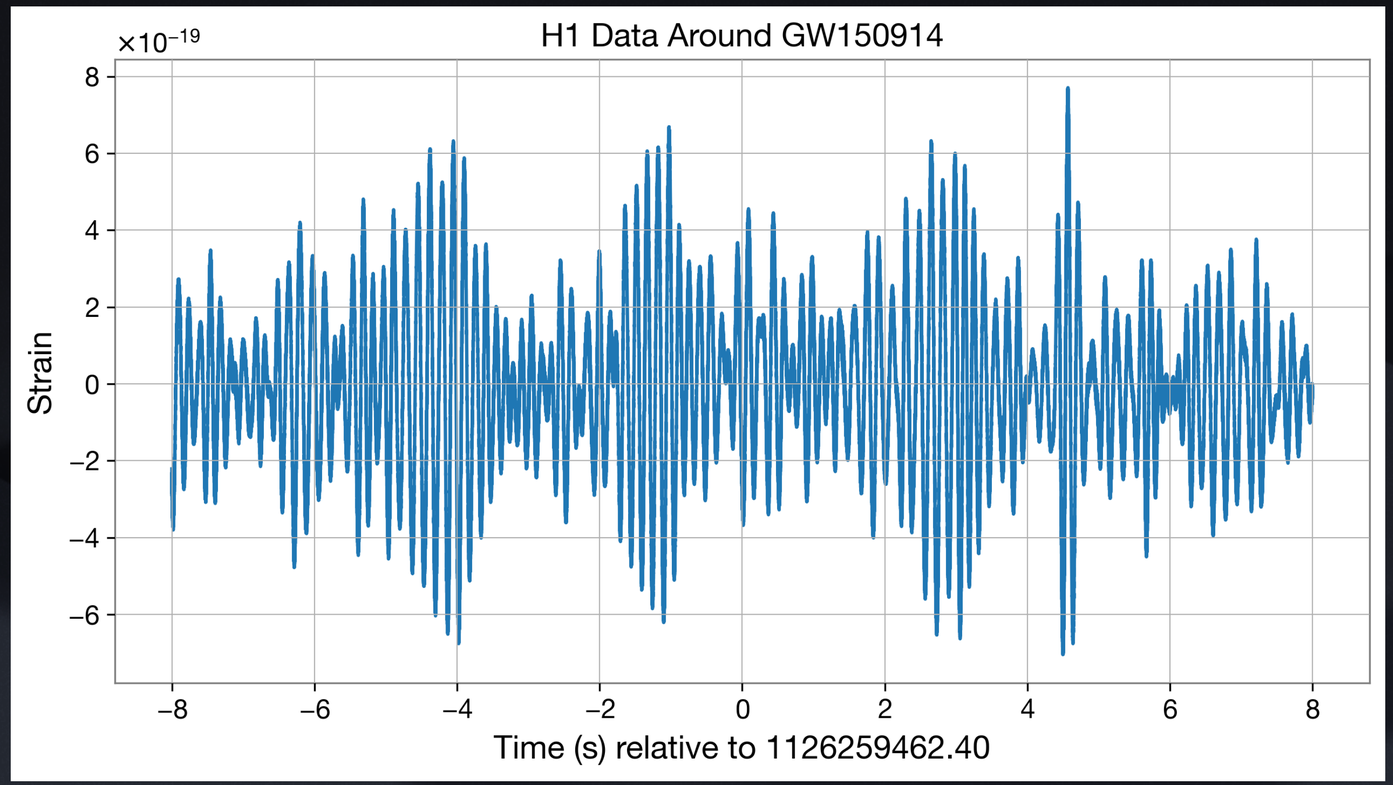
Yu-Chiung Lin
2026 Gravitational Wave Open Data Workshop in Taiwan
Taipei Municipal Chenggong High School
2026. 4. 18

Outline

- Basic data processing in the frequency domain
- Searching CBC signals using matched filtering
- Things we can do after the detection

Strain data from GW detector

- In this lecture, we focus on looking for signals from compact binary coalescence (CBC).
- Most of the time, what you see in the strain data is just noise.
- GW signals are usually invisible in the original strain data.



From time domain to frequency domain

- To understand the data characteristics better, we can decompose the periodic time series data into a Fourier series:

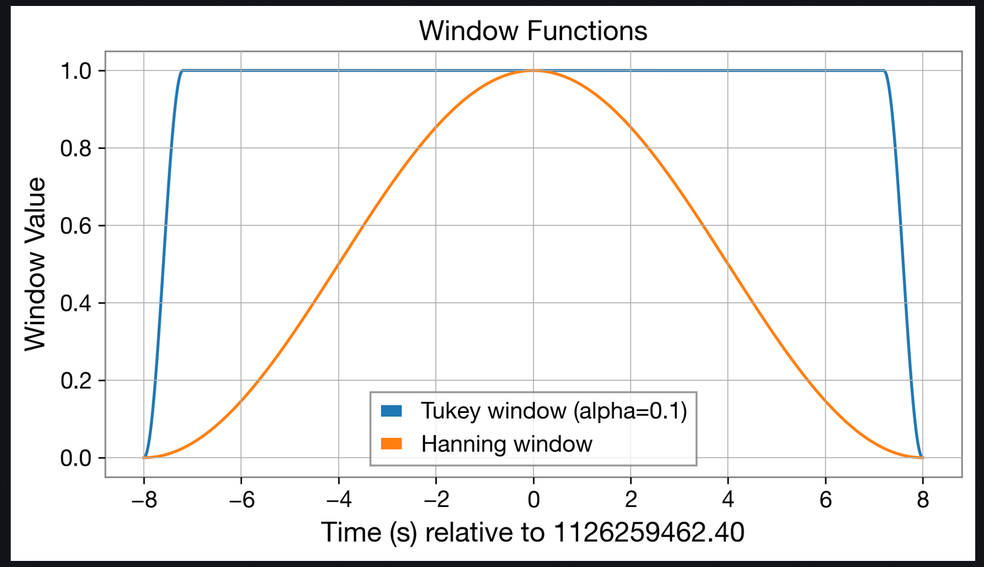
$$d(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left[a_n \cos(nt) + b_n \sin(nt) \right]$$

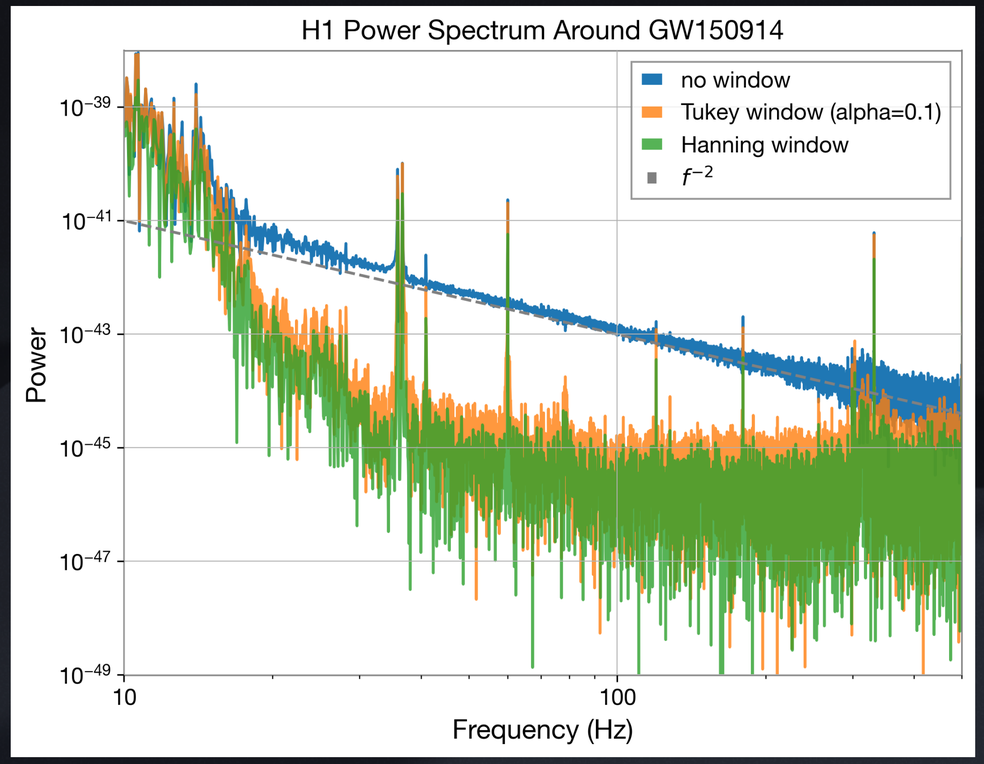
- Which is a combination of sine and cosine waves with different frequencies and amplitudes
- The Fourier series can be extended to the Fourier transform:

$$\tilde{d}(f) = \mathcal{F}(d(t)) = \int_{-\infty}^{\infty} d(t) e^{-i2\pi ft} dt$$

Window is important

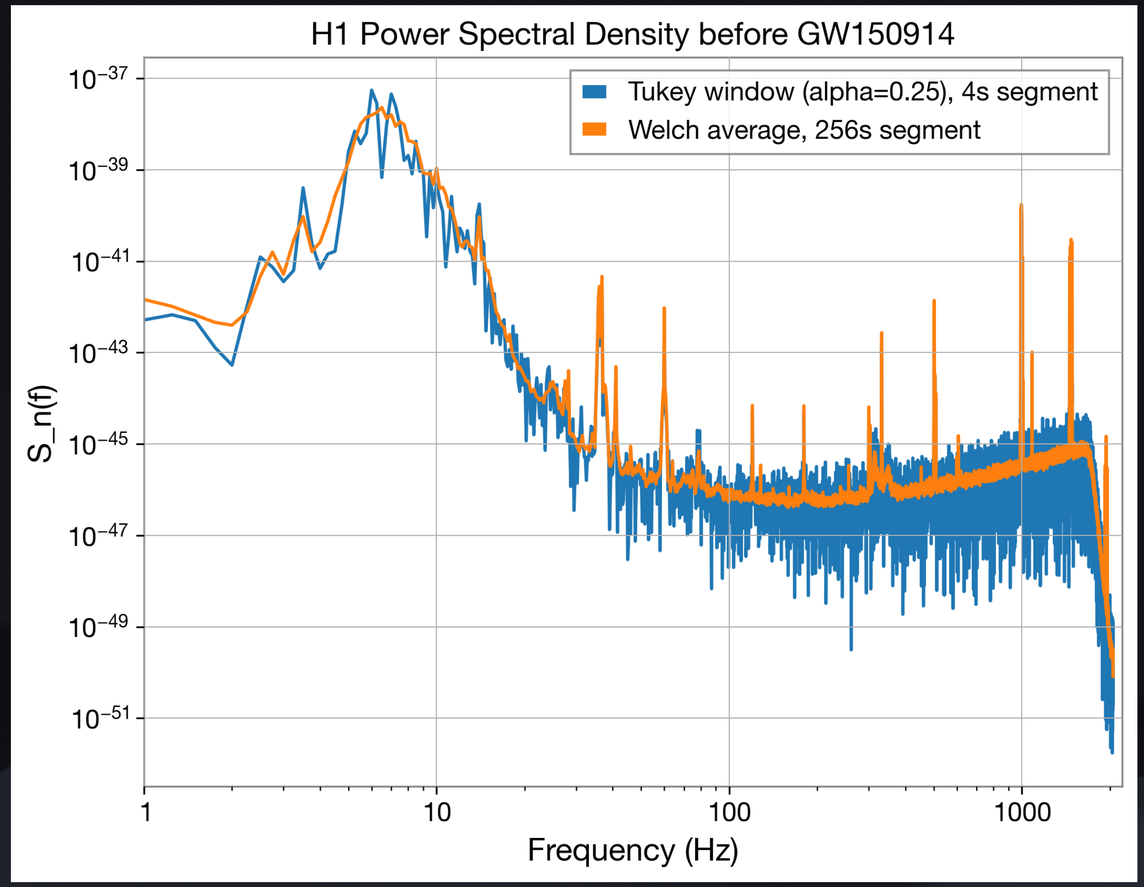
- Using the fast Fourier transform (FFT), we can easily obtain $\tilde{d}(f)$ and the power spectrum $|\tilde{d}(f)|^2$
- However, since the Fourier transform assumes the data to be periodic, a direct transform on non-periodic data can lead to spectral leakage
- Therefore, we have to apply a window function to make the data periodic
  - Caveat: applying a window function means some information will be lost, and therefore, reduces the signal power.





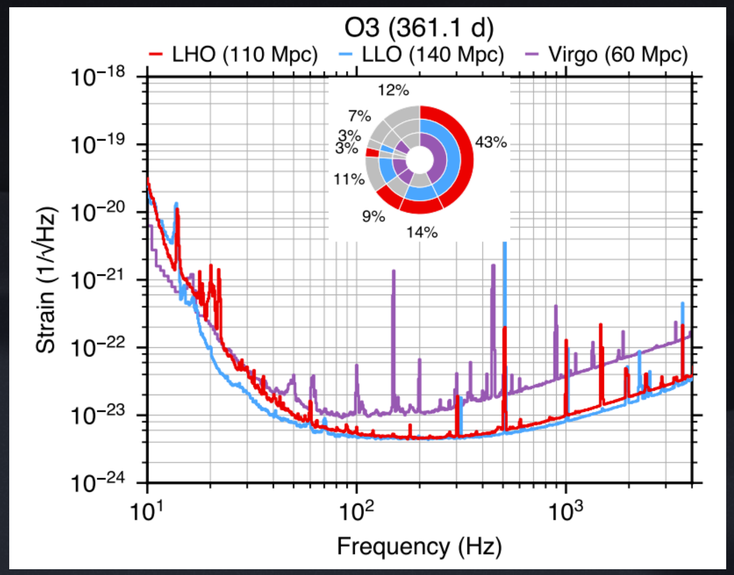
Background noise estimation

- From the power spectrum, we can calculate the (instantaneous) power spectral density (PSD): $S(f) = 2\Delta f\,|\,\tilde{d}(f)\,|^2$
- However, the PSD from a single window has high variance (and sometimes contains signals and glitches), therefore cannot represent the actual noise background
- Welch method: take the power spectrum from multiple windows with overlap, then take the average or median
  - P.S. Most of the packages have compensated for the power loss due to the window function



Detector Noise Changes Over Time

- Different detectors have different configurations and conditions; they have different noise backgrounds
- The background noise in the same detector can also change from time to time
  - The detector noise is non-stationary



Abec et al. 2025

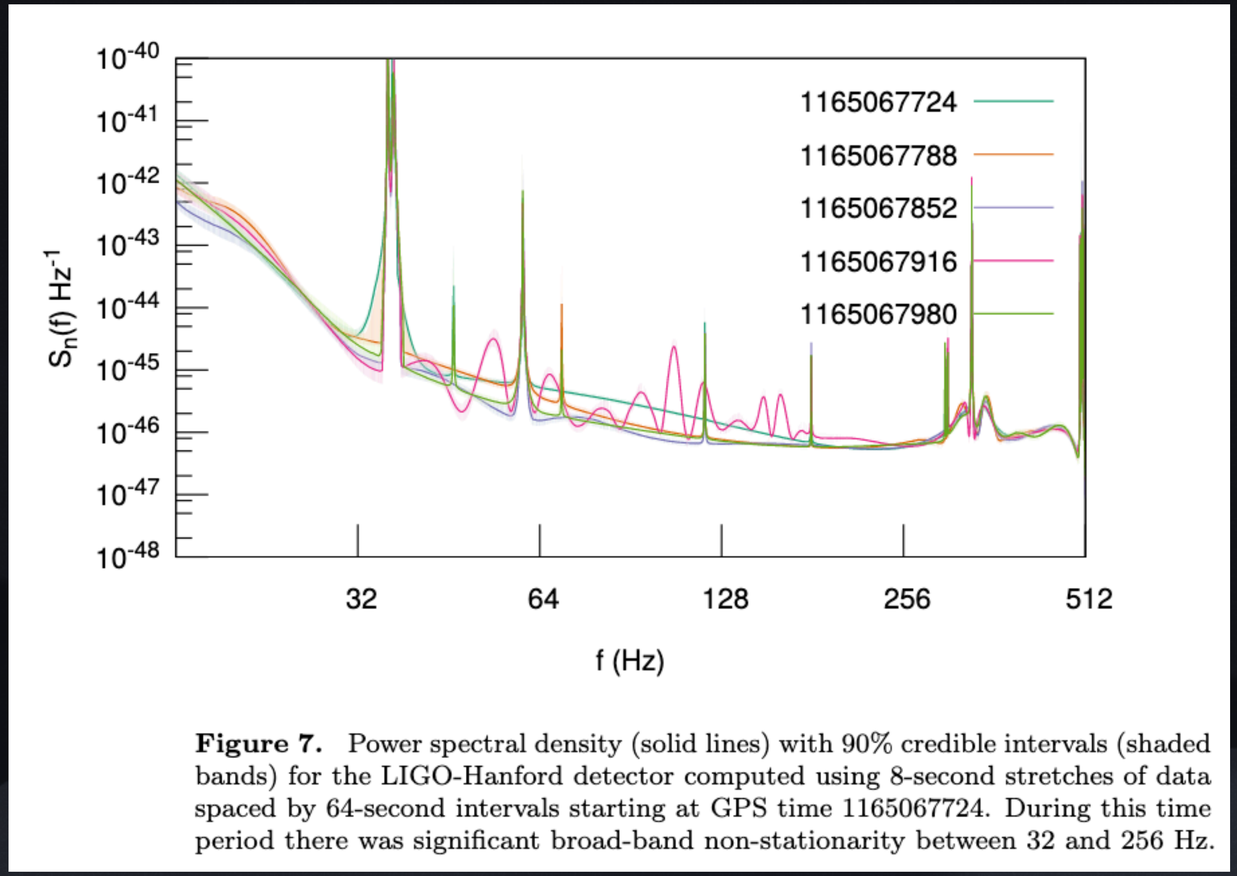
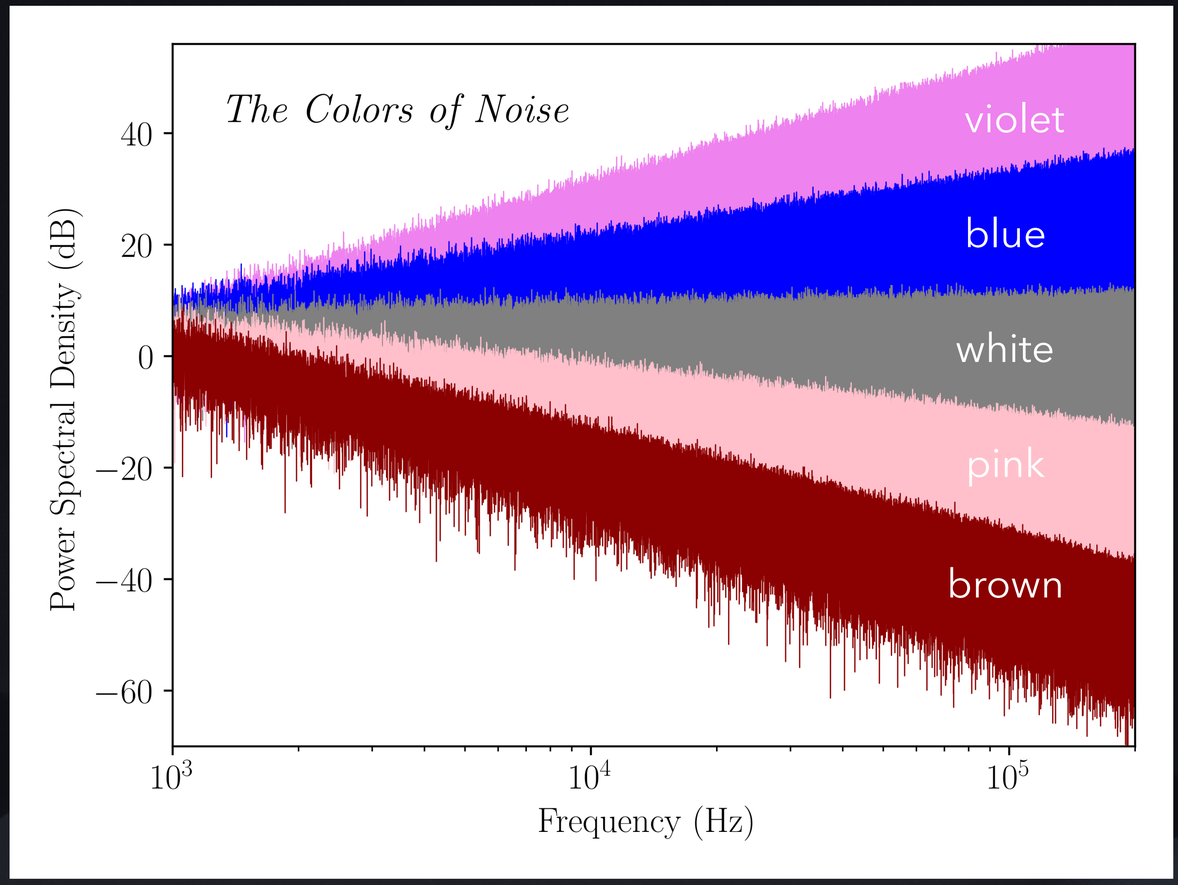


Figure 7. Power spectral density (solid lines) with 90% credible intervals (shaded bands) for the LIGO-Hanford detector computed using 8-second stretches of data spaced by 64-second intervals starting at GPS time 1165067724. During this time period there was significant broad-band non-stationarity between 32 and 256 Hz.

Abott et al. 2020

The Colors of Noise

- "White" noise: the random noise with a flat power spectrum
  - Has equal power in any band of given bandwidth
- Detector noises are "colored" noises
- Special colors of noise: the power spectrum proportional to $f^{-\beta}$
  - "Violet": $\beta = -2$
  - "blue": $\beta = -1$
  - "Pink": $\beta = 1$
  - "Brown": $\beta = 2$
- "Grey" noise: equal loudness at each frequency



Suppress the noise

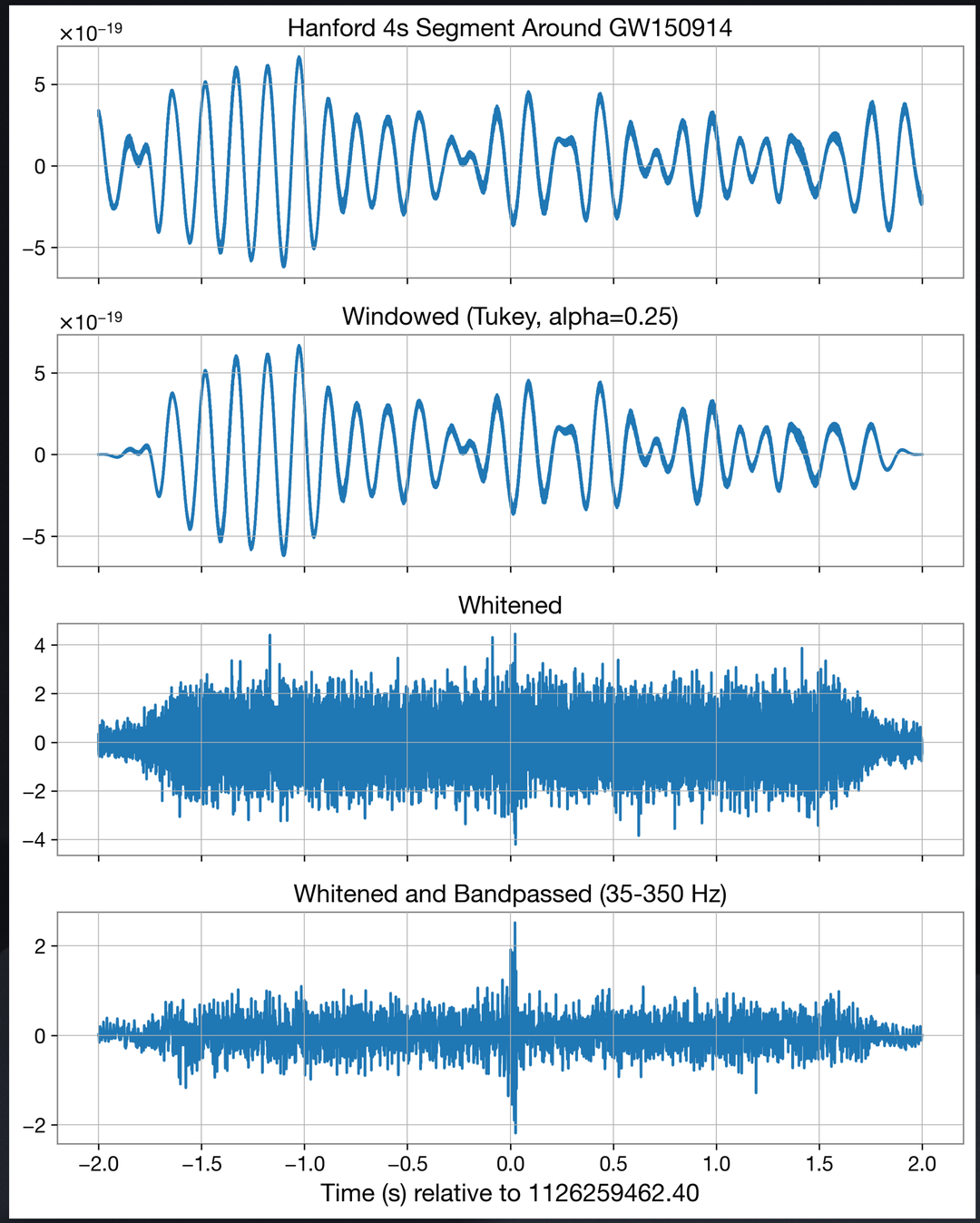
- With the PSD, we can suppress the noise contribution by whitening the data

$$\tilde{d}_w(f) = \frac{\tilde{d}(f)}{\sqrt{S_n(f)}}$$

- The whitened strain data can be obtained by

$$d_w(t) = \text{IFFT}(\tilde{d}_w(f))$$

- After band-passing, we can see some spikes around the event time
  - How do we know this is a GW signal?

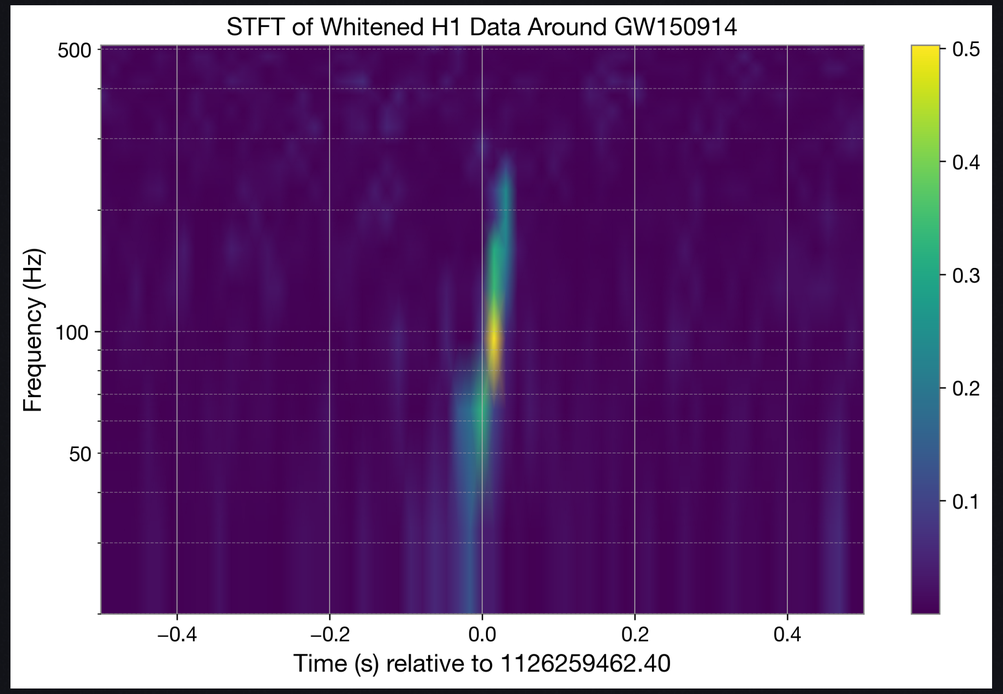


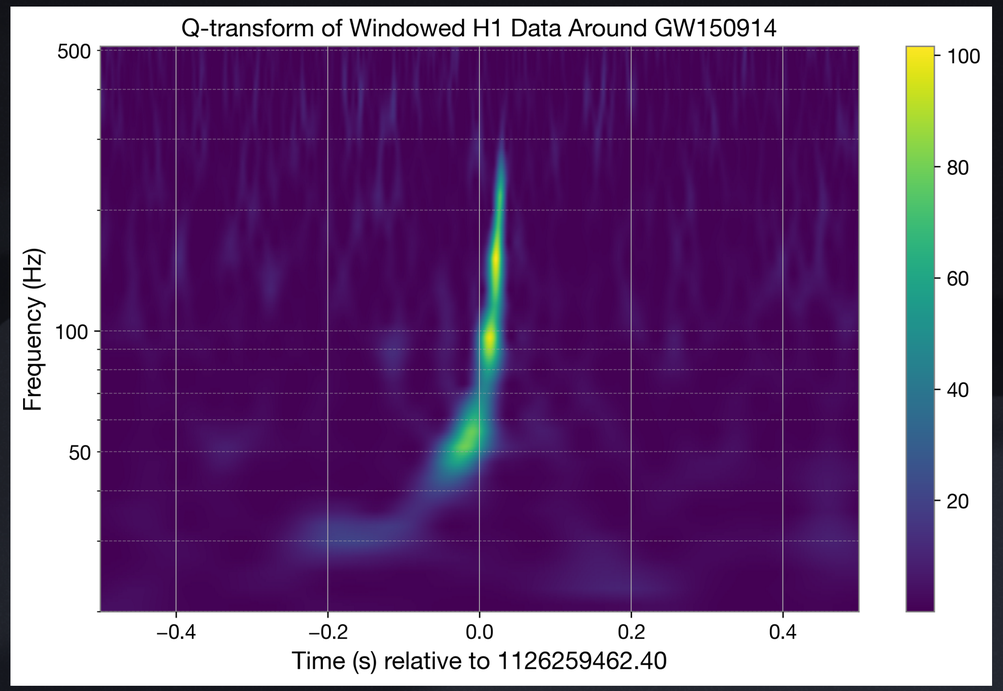
Q-Transform

- We can transform the whitened data into a spectrogram to see their behavior in both the time and frequency domains
- Approach 1: short-time Fourier transform (STFT)
  - Equal time and frequency bins
  - Hard to capture both low and high frequency features
- Approach 2: Q-transform
  - Different time and frequency resolutions at different frequency bins (controlled by Q factor)

$$Q = \frac{\delta f_k}{f_k}$$

  - Can capture all features at different frequencies



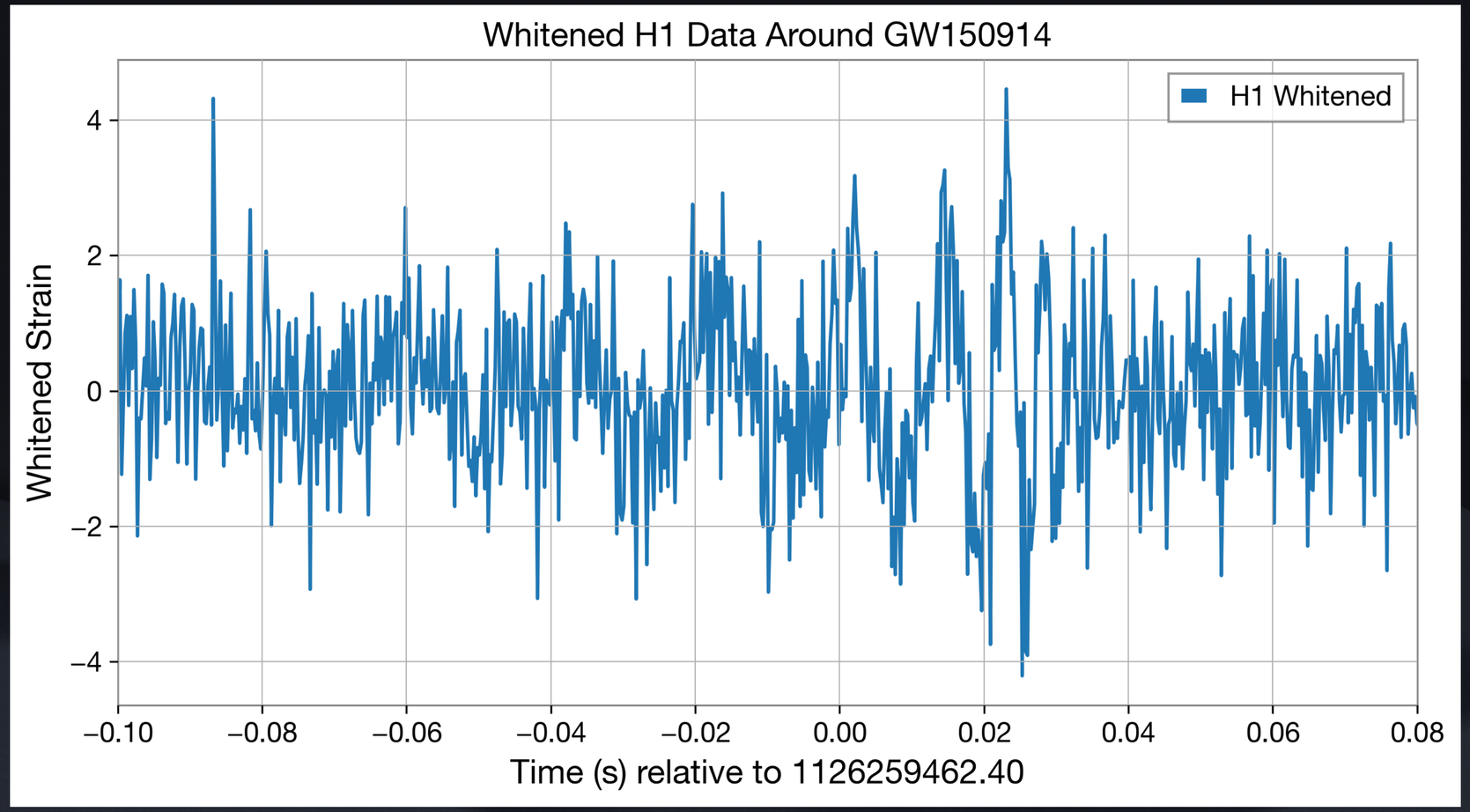


Outline

- Basic data processing in the frequency domain
- Searching CBC signals using matched filtering
- Things we can do after the detection

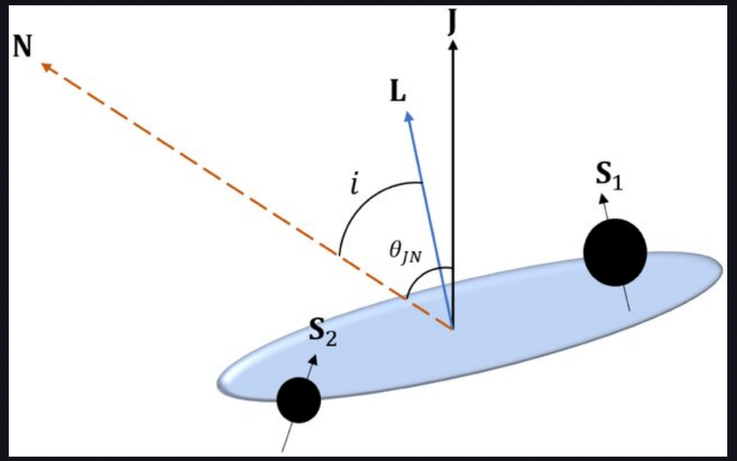
GW Signals in Time Domain

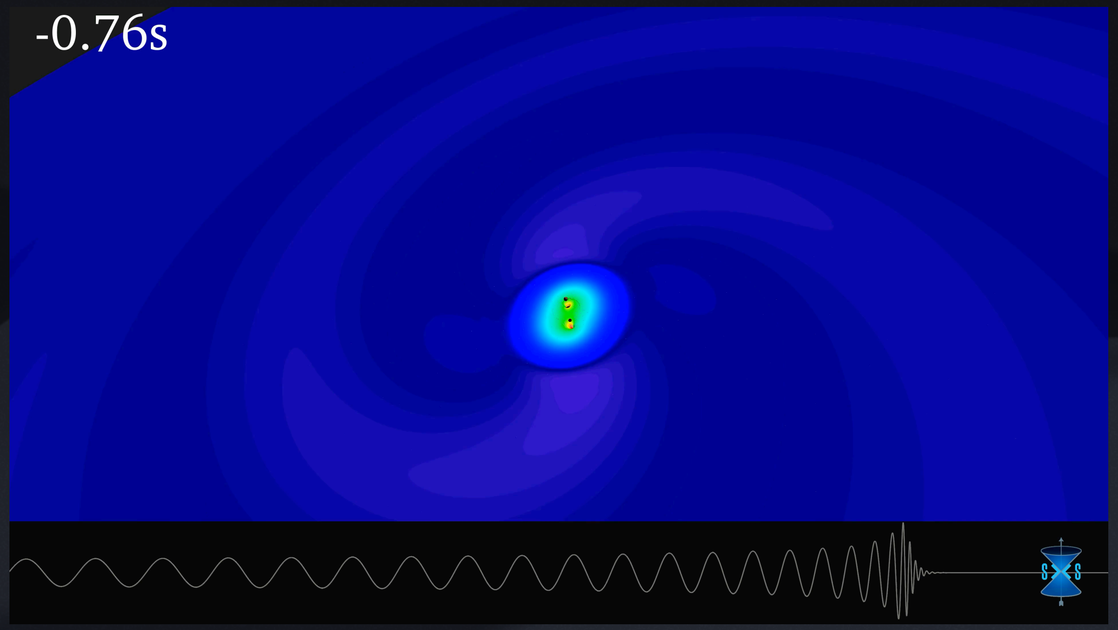
- It is unrealistic to watch the strain data with the naked eye during searches. A systematic method is necessary.



Modeling CBC Waveforms

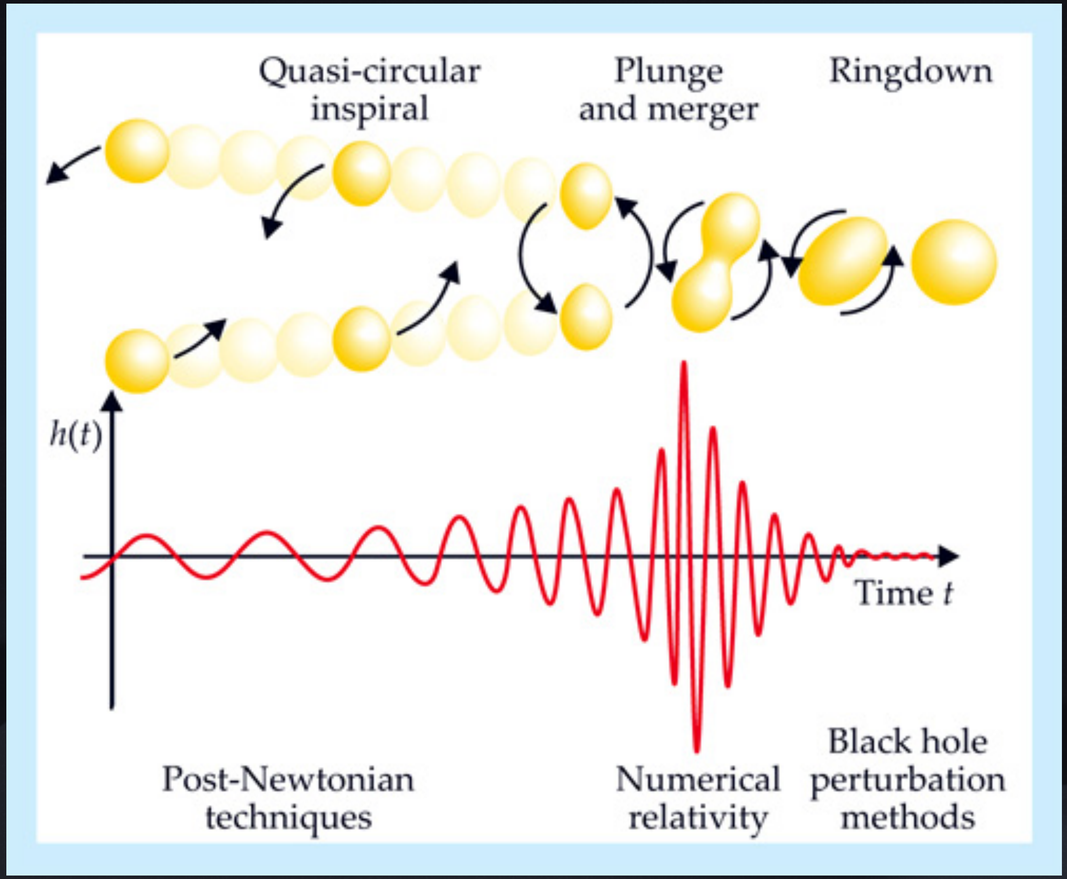
- GWs from CBC can be described by (at least) 15 parameters:
  - Intrinsic parameters (Object's own properties)
    - Component masses (2): m_1, m_2
    - Spin vectors (6): $\vec{S}_1, \vec{S}_2$
  - Extrinsic parameters (Relative properties between object and observer)
    - Sky location (2): (α, δ)
    - Luminosity distance (1): D_L
    - Binary orientation (1): θ_{JN}
    - Polarization angle (1): ψ
    - Merger phase (1): φ
    - Merger time (1): t_c
- More parameters may be required by complex models
  - Higher modes, tidal forces, eccentricity, etc.





CBC Waveform Simulation

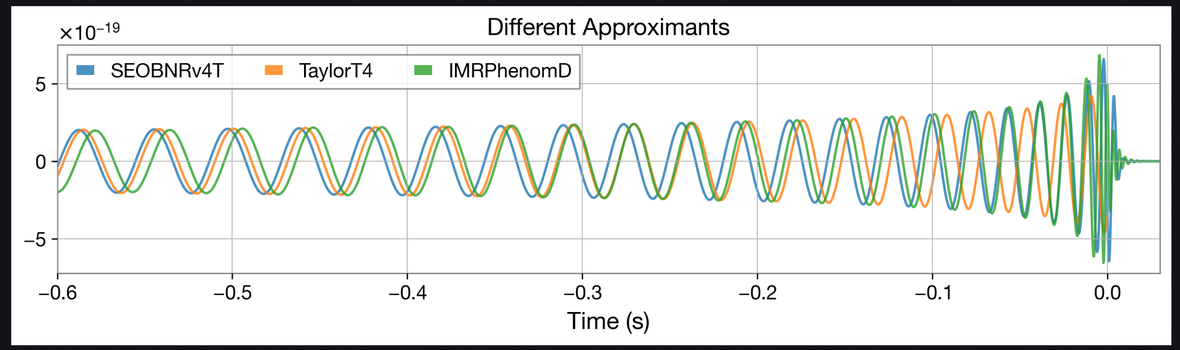
- Compact binary coalescence consists of three stages: inspiral, merger, and ringdown
- In the inspiral stage, we can use the post-Newtonian techniques to describe and model the waveform
- The merger stage is a complex process and can only be solved using numerical relativity (NR) simulations
- The ringdown process can be described using perturbation methods
- We have developed several approximants to efficiently generate the waveform

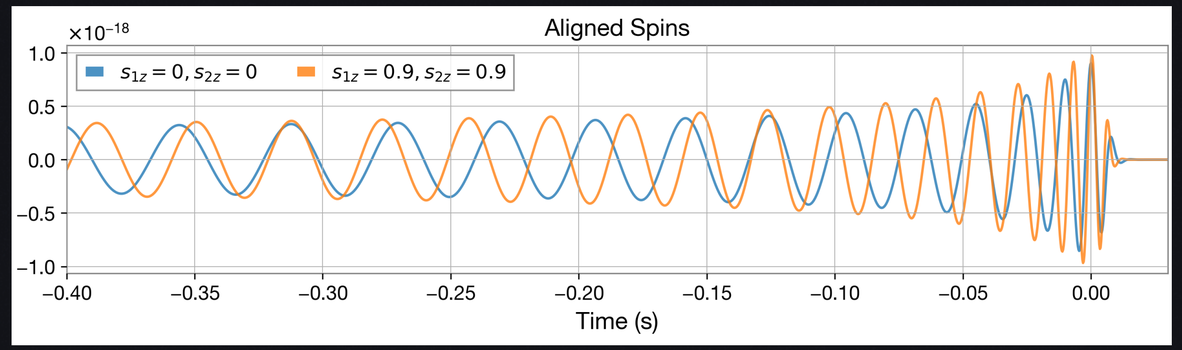


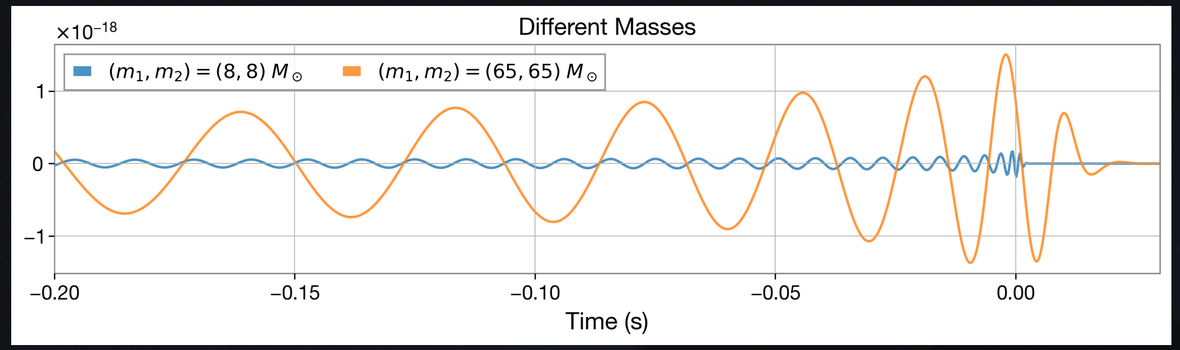
(Major) Families of CBC approximants

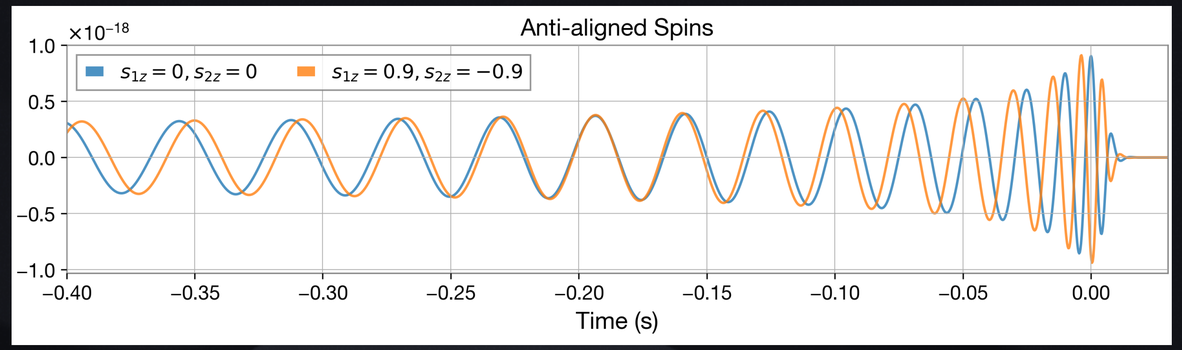
Taylor	Effective one-body (EOB)	IMR (inspiral-merger-ringdown) Phenomenological	NR Surrogate
TaylorT1, TaylorT2, TaylorT3, TaylorF1,TaylorF2, TaylorF2Ecc, TaylorF2NLTides,TaylorR2F4, TaylorF2RedSpin,TaylorF2RedSpinTidal, SpinTaylorT1, SpinTaylorT2, SpinTaylorT3, SpinTaylorT4, SpinTaylorT5, SpinTaylorF2, SpinTaylorFrameless, SpinTaylor, TaylorEt,TaylorT4,TaylorN, …	EOB, EOBNR, EOBNRv2, EOBNRv2HM, EOBNRv2_ROM,EOBNRv2HM_ROM, TEOBResum_ROM, SEOBNRv1, SEOBNRv2,SEOBNRv2_opt , SEOBNRv3 , SEOBNRv3_pert , SEOBNRv3_opt ,SEOBNRv3_opt_rk4 , SEOBNRv4 , SEOBNRv4_opt , SEOBNRv4P ,SEOBNRv4PHM, SEOBNRv2T, SEOBNRv4T, SEOBNRv1_ROM_EffectiveSpin, …	IMRPhenomA,IMRPhenomB, IMRPhenomFA,IMRPhenomFB, IMRPhenomC, IMRPhenomD, IMRPhenomD_NRTidal, IMRPhenomD_NRTidalv2,IMRPhenomNSBH, IMRPhenomHM, IMRPhenomP, IMRPhenomPv2,IMRPhenomPv2_NRTidal,IMRPhenomPv2_NRTidalv2, IMRPhenomFC, …	NRSur4d2s, NRSur7dq2, NRSur7dq4, NRHybSur3dq8, …
Post-Newtonian (PN) approximants with Taylor expansion	Effective one-body approach to solve two-body problem in GR	PN + other analytical functions to phenomenologically describe NR waveforms	Surrogate models directly built from NR simulations

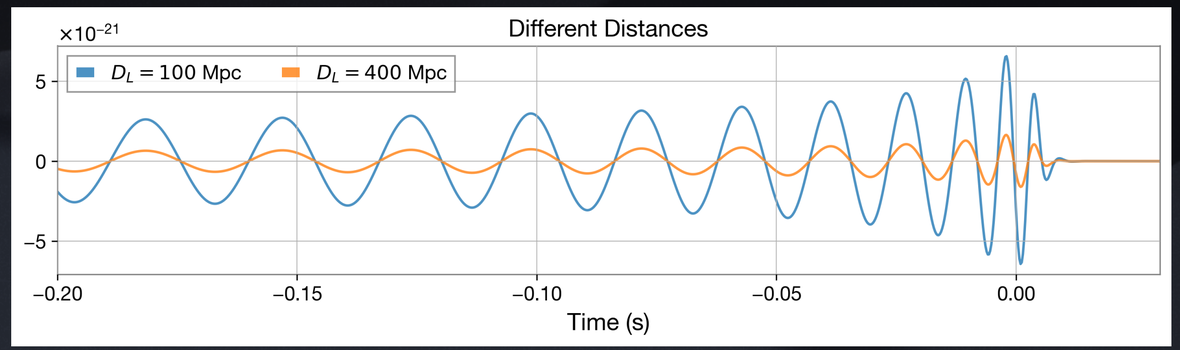
Phenomenology of CBC signals

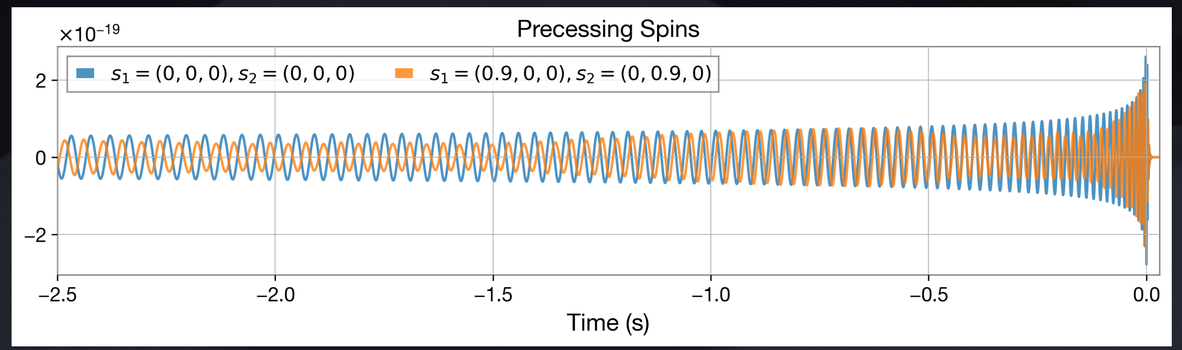












Searching GW Signals in the Strain Data

- We can start by assuming the strain data is a combination of Gaussian noise $n(t)$ and an unknown GW signal $\eta(t)$:

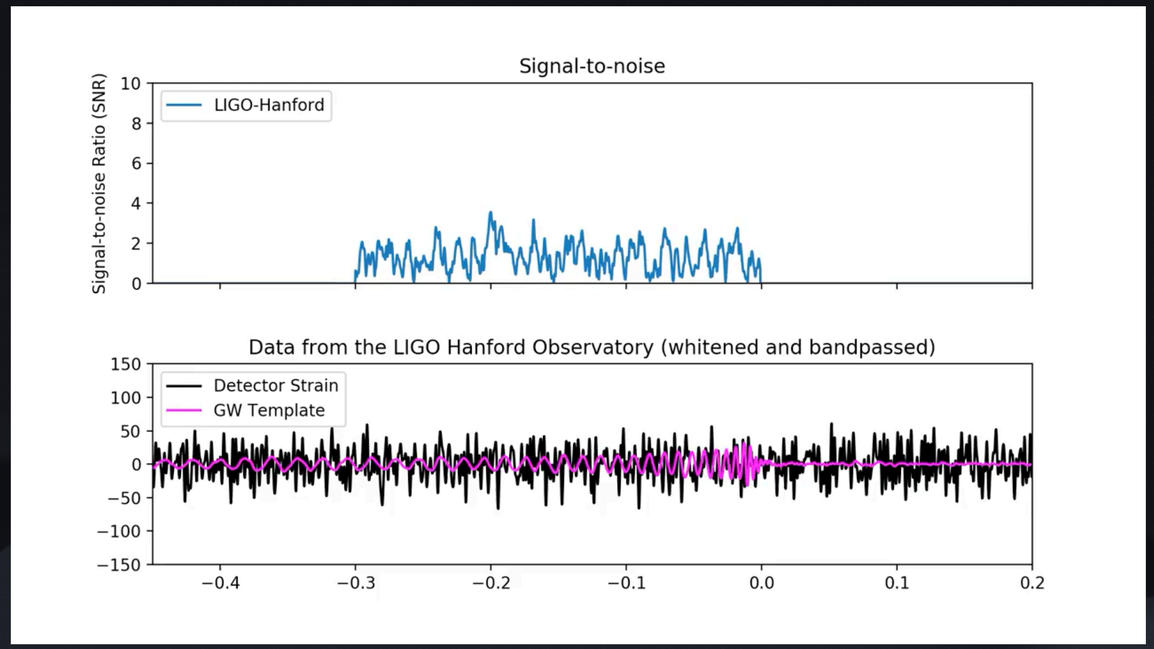
$$d(t) = n(t) + \eta(t)$$

- Given a known template $h(\theta)$ with a set of parameters θ, we can calculate the likelihood ratio between the signal hypothesis and the null hypothesis:

$$\log \Lambda(d\,|\,\theta) = \frac{H_1}{H_0} = (d\,|\,h(\theta)) - \frac{1}{2}(h(\theta)\,|\,h(\theta))$$

- The quantity $(d\,|\,h(\theta))$ is the matched filter:

$$(d\,|\,h) = 4\int_0^\infty \frac{\tilde{d}(f)\tilde{h}^*(f)}{S_n(f)}df$$



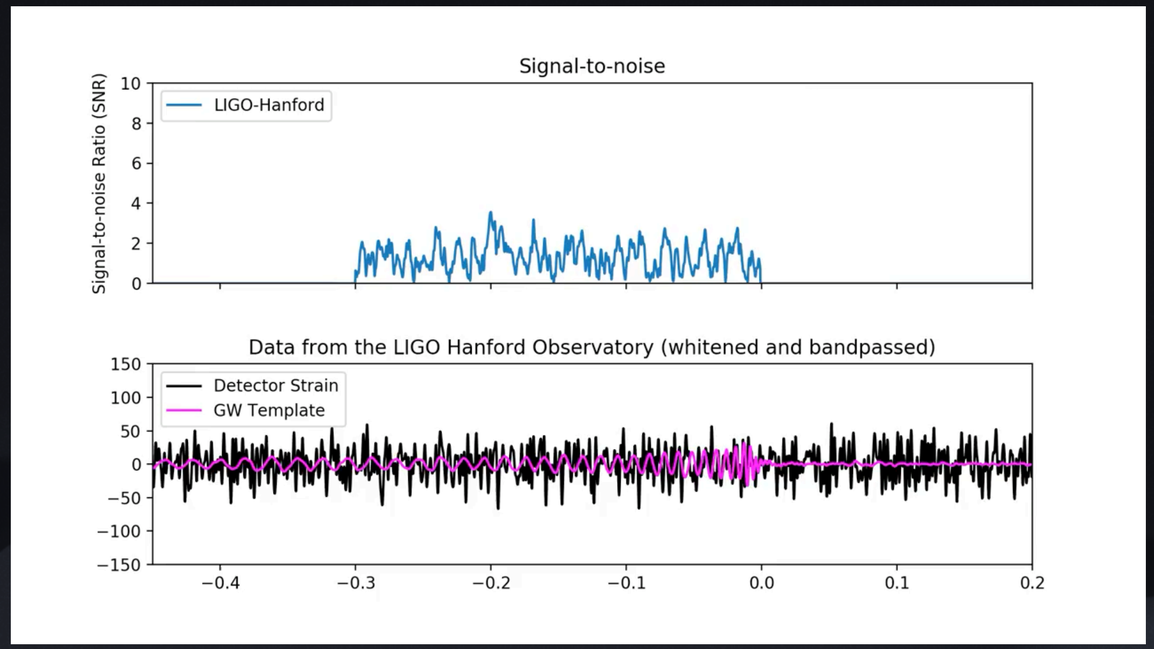
Searching GW Signals in the Strain Data

- The matched filter is an optimal statistic in the Gaussian noise
- However, the parameter set θ contains both extrinsic parameters, like amplitude A, phase ϕ, and arrival time t, and intrinsic parameters μ.
- We can rewrite the template $h(\theta)$ as

$$h(\theta) = A\hat{h}_+(t,\mu)\cos\phi + A\hat{h}_\times(t,\mu)\sin\phi$$

- $\hat{h}_+$ and $\hat{h}_\times$ are normalized waveforms with plus and cross polarizations:

$$\hat{h}_+ = \frac{h_+}{\sqrt{(h_+|h_+)}}$$

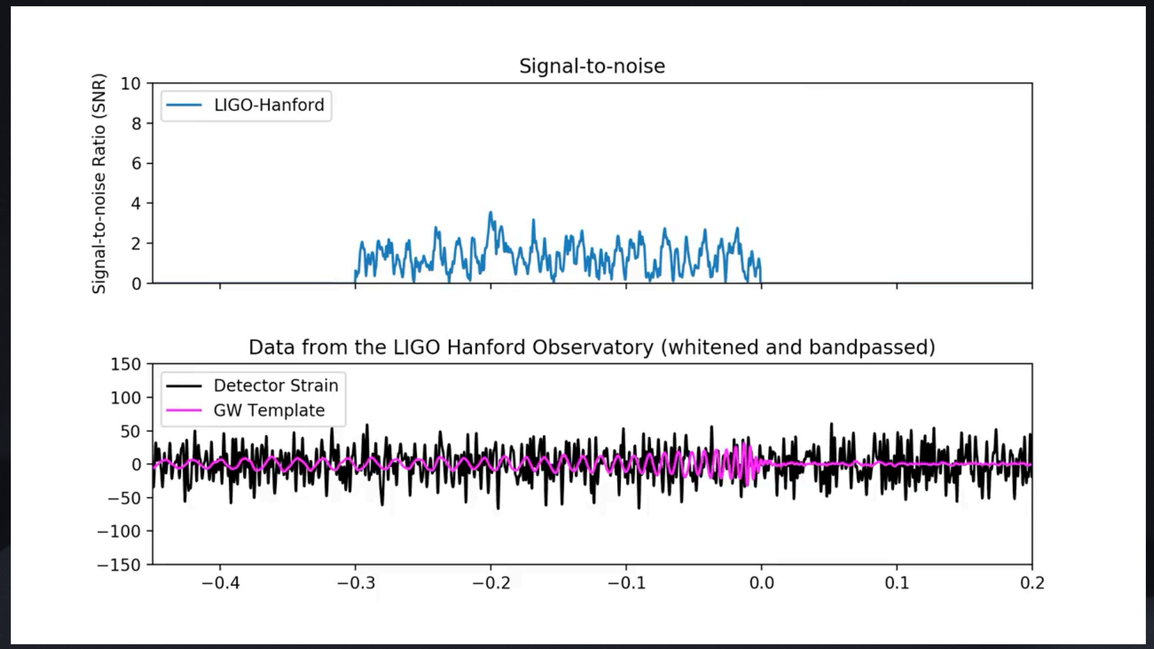


Searching GW Signals in the Strain Data

- After some derivations, we can obtain the matched-filter signal-to-noise ratio (SNR) time series for a normalized template $\hat{h}_+$:

$$z(t,\mu) = 4\int_0^\infty \frac{\tilde{d}(f)\tilde{\hat{h}}_+^*(f,\mu)}{S_n(f)} e^{2\pi i f t} df$$

- We can know the amplitude $|z|$ and phase $\arg(z)$ from the complex SNR time series
- The arrival time t can be known by finding the time of maximum amplitude
- The spikes of the SNR series are called triggers

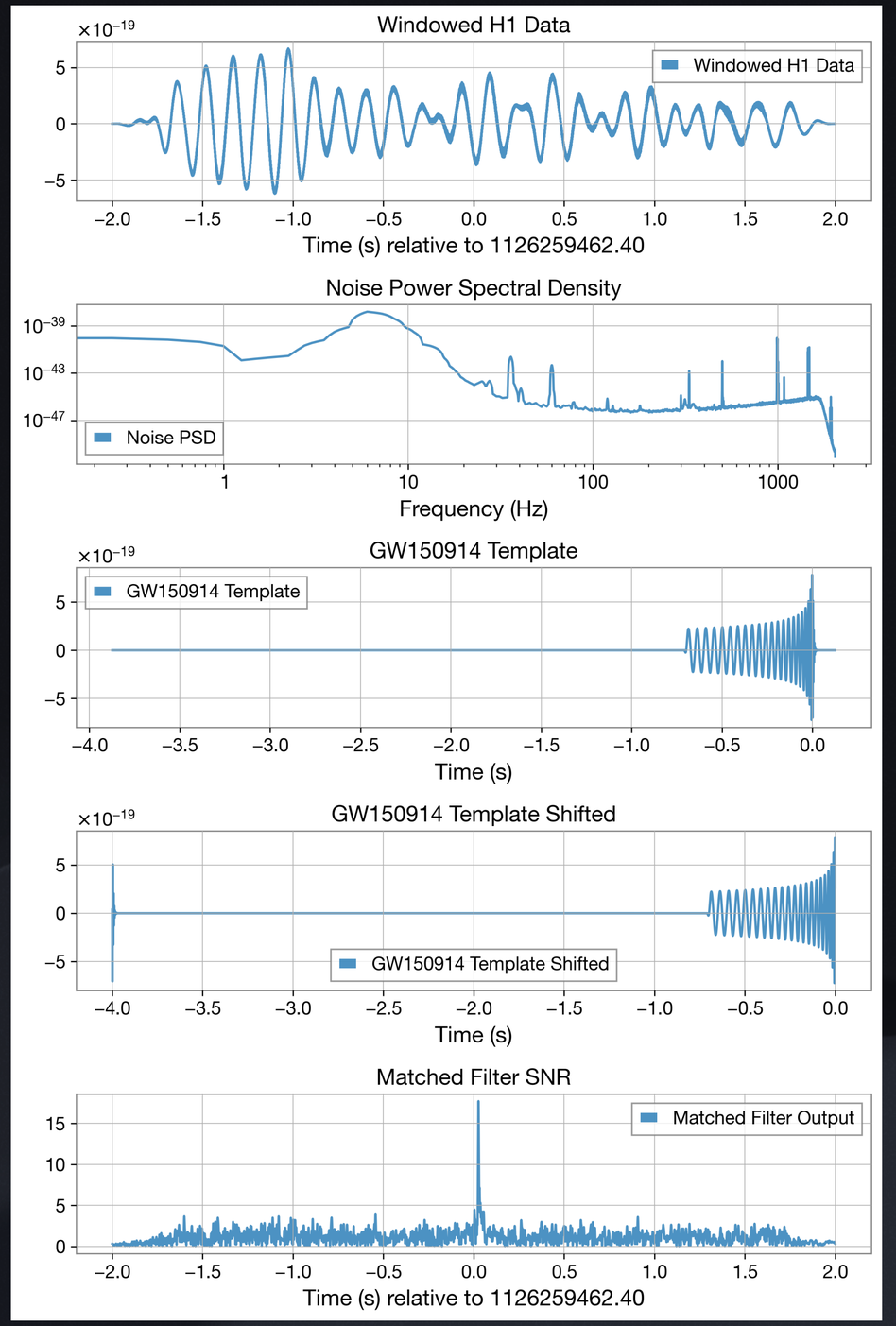


How to Matched Filtering

1. Select the strain time series $d(t)$ we are interested in
2. Calculate the PSD from the previous time segment
3. Generate an in-phase GW waveform $h_+(t)$, taper the edge, and resize to match the data length
4. Cyclic-shift the waveform so that the merger occurs at the end of the time series (so that we can match the arrival time with the merger time)
   - ✳ You can also generate a frequency-domain waveform to skip 3. and 4.
5. Perform matched filtering

$$z(t) = \frac{(d\,|\,h)}{\sqrt{(h\,|\,h)}}$$

6. Find the peak amplitude $\max(|z|)$ and its time



Matched Filter and Cross-Correlation

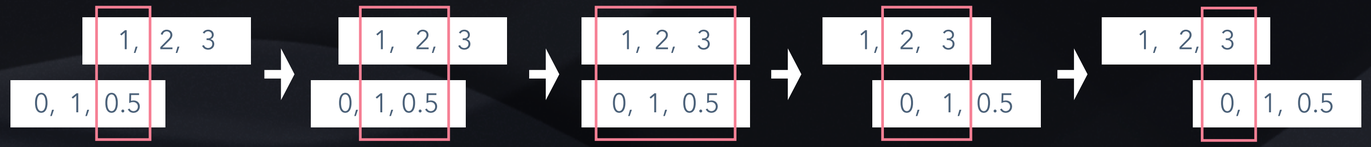
- The multiplication in the Fourier domain is equal to the cross-correlation in the time domain:

$$\mathscr{F}((F * G)(t)) = \mathscr{F}(F(t)) \cdot \mathscr{F}(G(t))$$

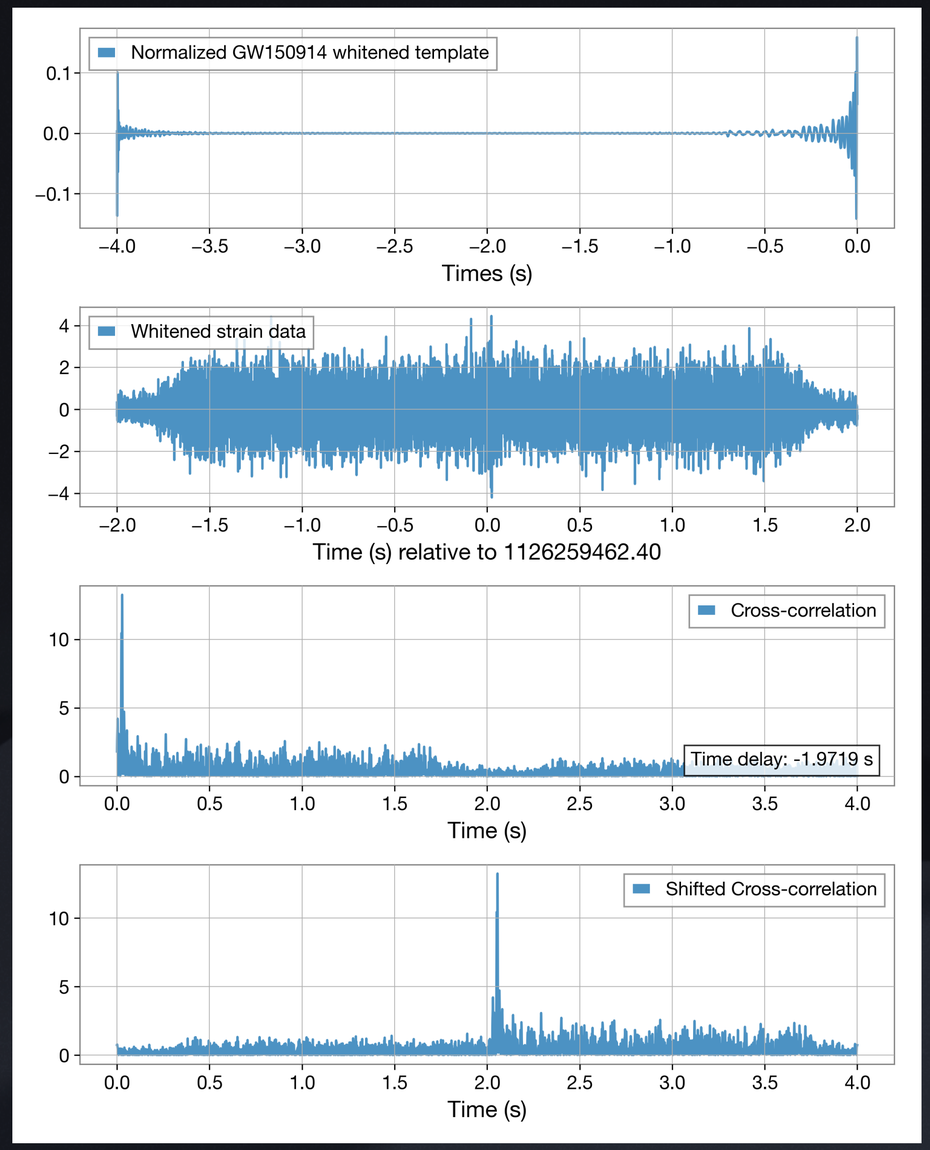
$$(F * G)(\tau) = \int_{-\infty}^{\infty} \overline{F(t-\tau)} G(t) dt$$

- Cross-correlation example:

```
>>> np.correlate([1, 2, 3], [0, 1, 0.5], "full")
array([0.5,  2. ,  3.5,  3. ,  0. ])
```

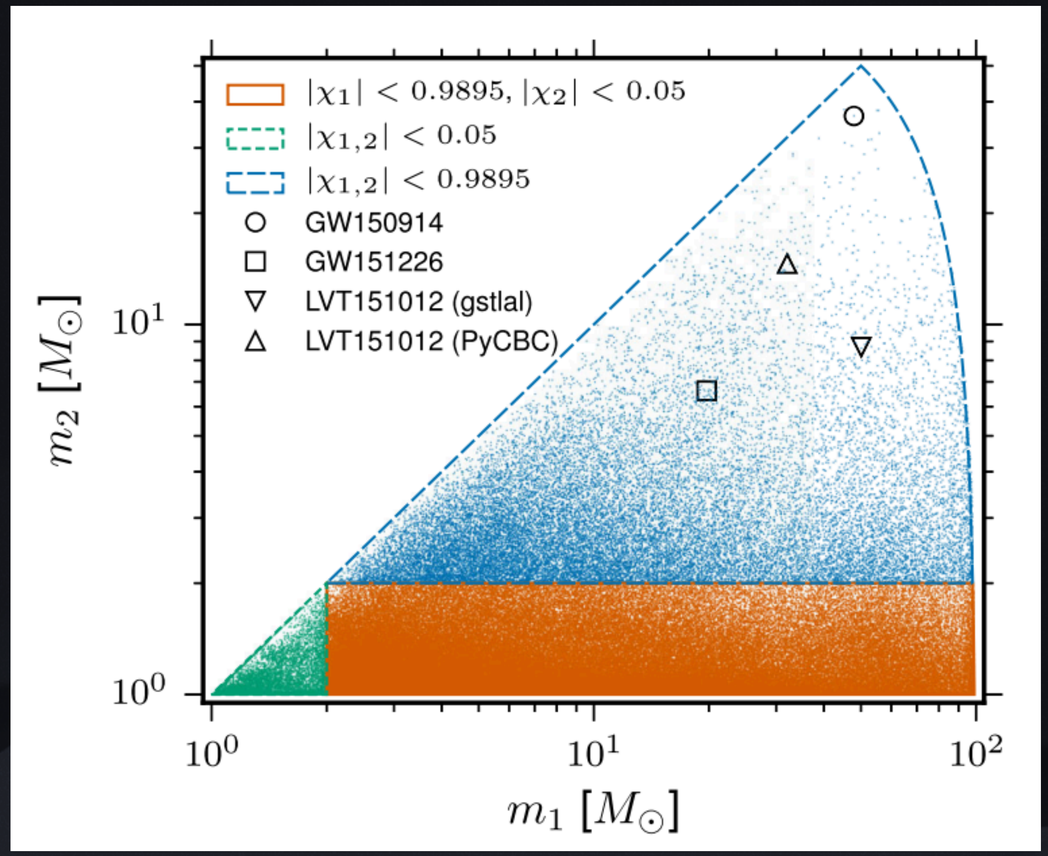


- The matched filtering process is equivalent to cross-correlating the whitened strain data with the whitened template



Template Bank

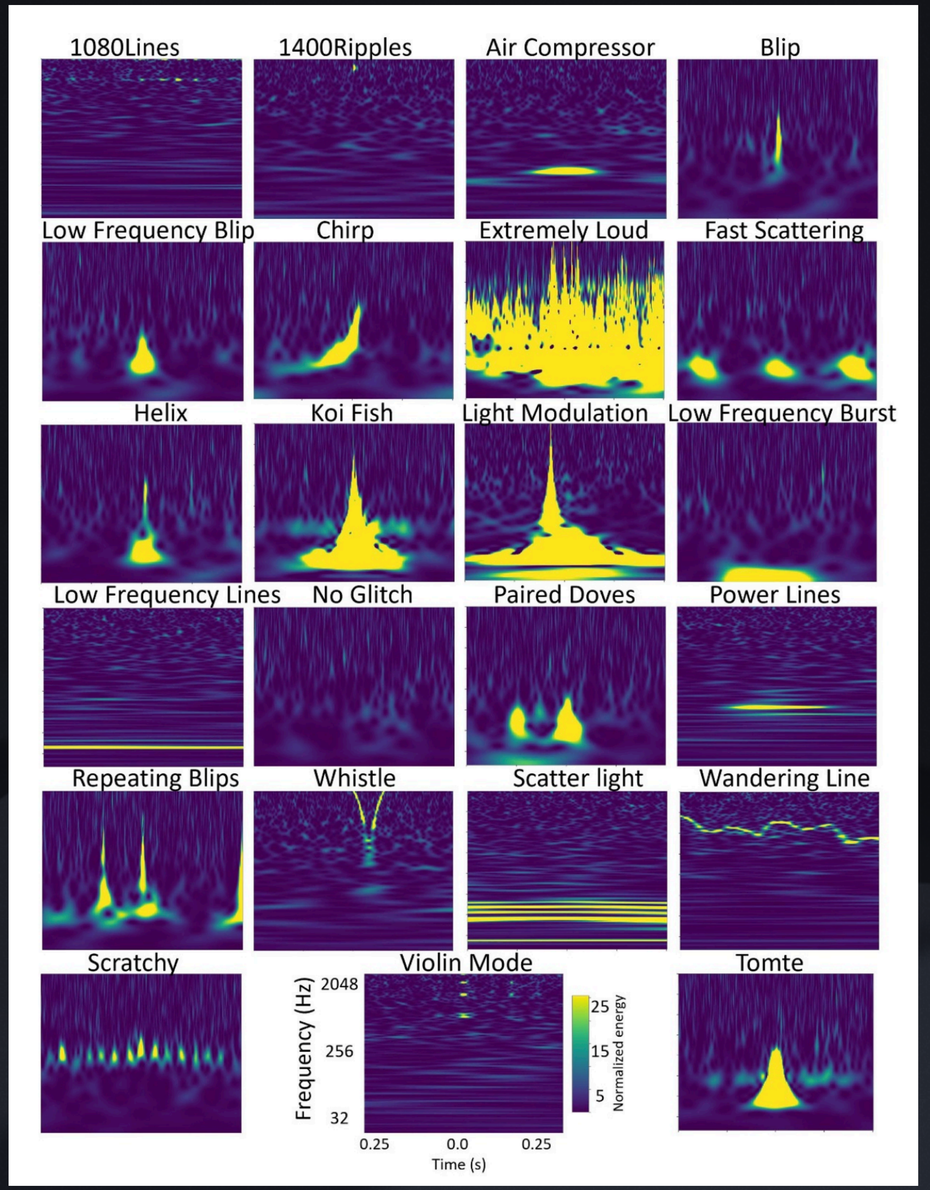
- In general, we do not know the existence or exact parameters of the GW signal in the strain data
- We have to match-filter the data with templates covering a large parameter space, including component masses and spins
- Find the balance: we need template parameters dense enough to capture the true signal, while maintaining the computational efficiency
  - Demand that the loss in SNR between the true signal and the best-fit template is less than 3%



Abott et al. 2016

Challenges of Matched Filtering

- Matched filtering SNR is an optimal statistic in Gaussian noise
- However, the strain data contains non-Gaussian glitches that can also give high SNR
- GravitySPY: a machine-learning project to identify glitches in the strain data
  - Using the “labor intelligence” to label glitches for training



Yunan Wu et al. 2025

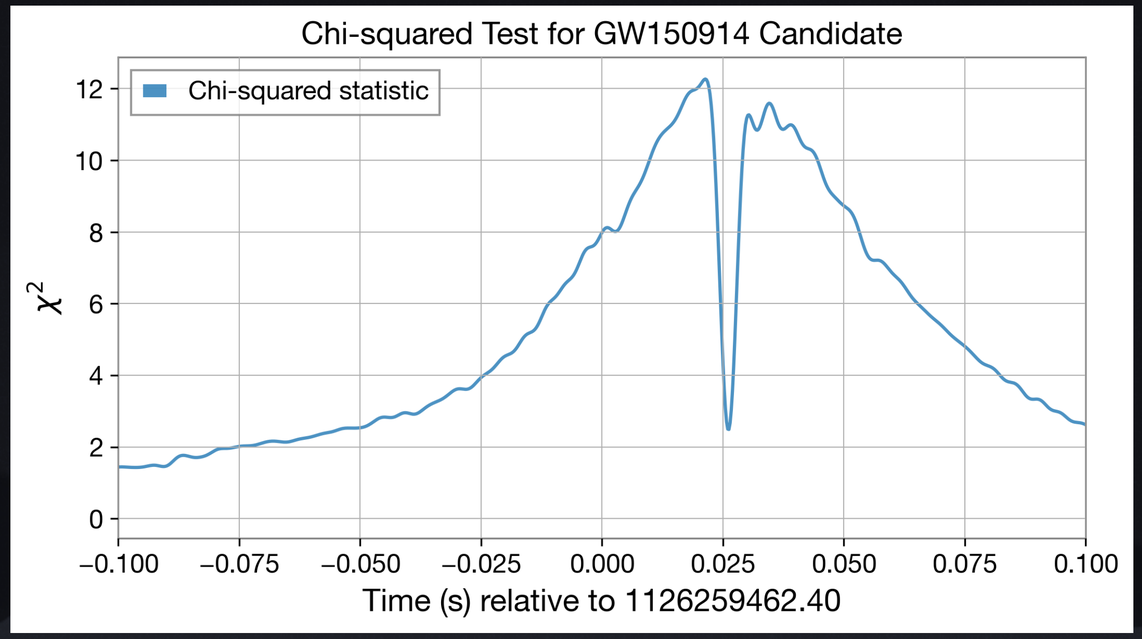
Signal Consistency Test

- We can use the χ^2 (chi-squared) test to discriminate signals and glitches
- χ^2 test divides the matched filtering into p frequency bands [B. Allen 2005]

$$\chi_r^2 = \frac{\chi^2}{2p-2} = \frac{p}{2p-2}\sum_{l=0}^{p}(z_l - \frac{z}{p})^2$$

  - The frequency bands are chosen so that the power of the template is equally distributed among these bands

- If the template matches the signal, χ_r^2 (reduced χ^2) will be close to 1
- χ_r^2 will be large if we find glitches in the data instead of signals



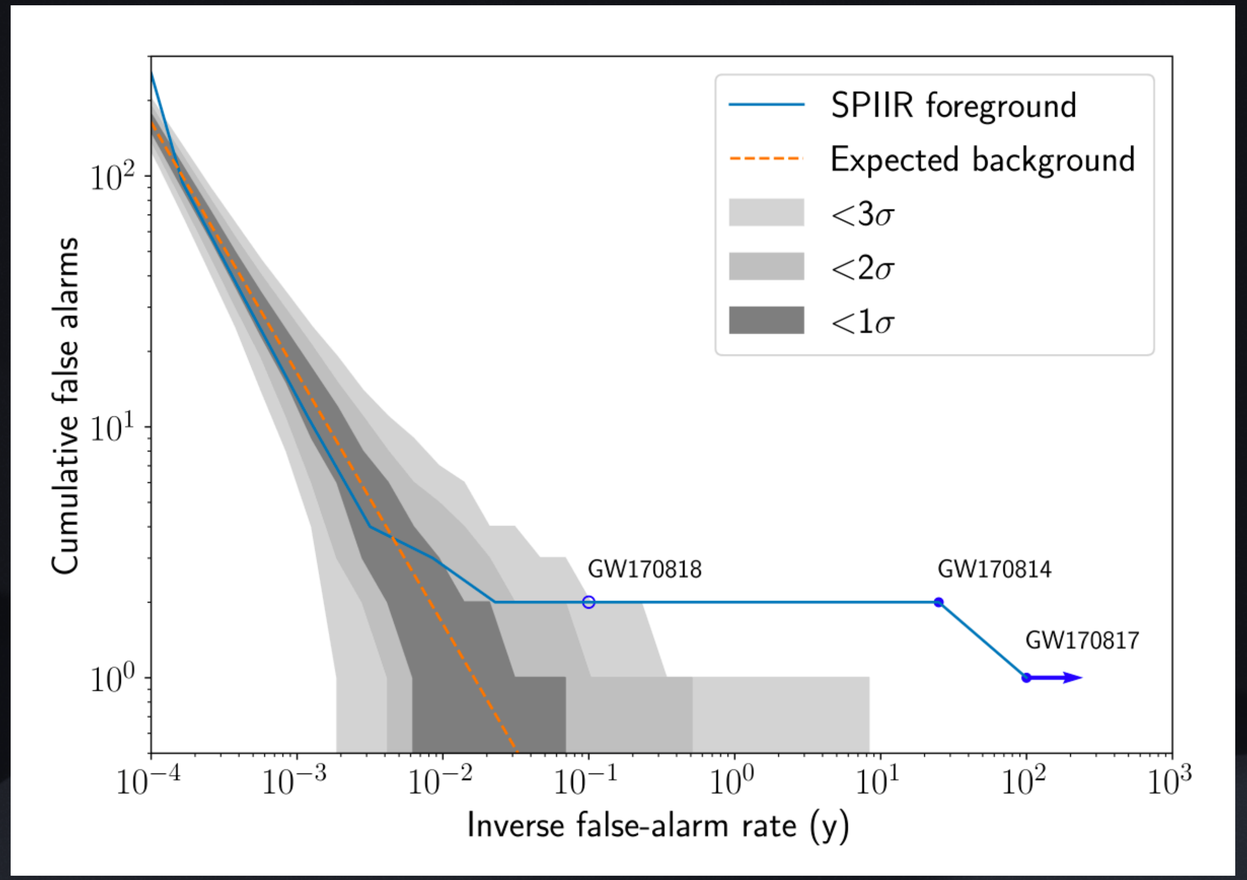
More Detectors, Less False Alarms

- **Coincidence test**: the time difference of triggers from two detectors should be less than the light-travel time between each detector
  - Light-travel time between Hanford and Livingston ~15ms
  - The masses and spins template parameters from triggers should also be the same
- **Coherence test**: not only checking the arrival times but also looking at the phases of the SNR time series
  - Signals detected in detectors should come from the same source (sky area)
  - Further reduce the possibility of coincidence glitches
- The trigger that passes these tests is called an event

False Alarm Rate

- We use the false alarm rate (FAR) to quantify the detection significance
  - How often can such a coincident/coherent trigger be produced by the noise fluctuations in the current noise background?
- Calculate FAR:
  - Collect background triggers by searching time-shifted strain data
  - For each trigger, compute the ranking statistics $\mathscr{R}$ based on the matched filtering SNR and signal consistency tests
  - For a newly detected trigger, calculate the false alarm probability (FAP)
  - We can obtain the FAR based on the FAP, counts of background triggers, and the total search time

Qi Chu et al. 2021



Current Search Pipelines

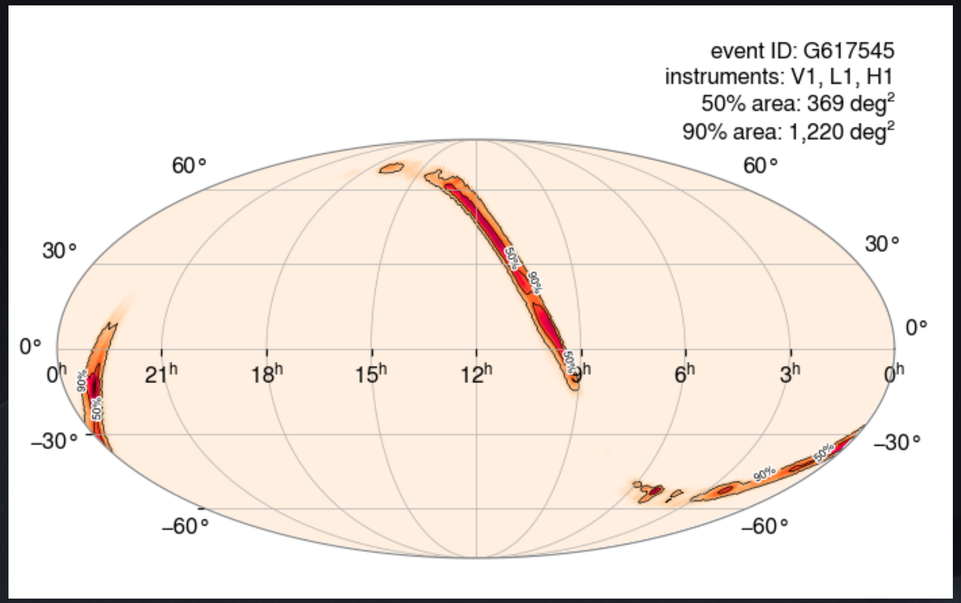
- Template-based searches (matched filtering)
  - GstLAL: https://lscsoft.docs.ligo.org/gstlal/
  - PyCBC: https://pycbc.org/pycbc/latest/html/index.html
  - MBTA: T. Adams et al. 2016 (https://arxiv.org/abs/1512.02864)
  - SPIIR: Qi Chu et al. 2021 (https://arxiv.org/abs/2011.06787)
- Unmodeled searches (looking for coherent excess power)
  - CWB (Coherent Wave Burst): https://gwburst.gitlab.io/
  - X-Pipeline: P. J. Sutton et al. 2010 (https://arxiv.org/abs/0908.3665)
- Machine-learning pipelines
  - Aframe: https://ml4gw-aframe.readthedocs.io/en/latest/index.html
  - Mly: V. Skliris et al. 2024 (https://arxiv.org/abs/2009.14611)

Outline

- Basic data processing in the frequency domain
- Searching CBC signals using matched filtering
- Things we can do after the detection

Sky Localization and Event Alerts

- Matched filter SNR time series contains time, amplitude, and phase information
- With the SNR time series from different detectors, we can rapidly localize the source of the GW signal using the BAYESTAR software.
- During the observation run, when a pipeline detects a GW event, it will send an alert to the Gravitational-Wave Candidate Event Database (GraceDB)
- The rapid-response team (RRT) will validate the event
- If the event is significant, an alert will be sent to the General Coordinates Network (GCN) for other optical or neutrino observatories to perform follow-up observations



GraceDB

Parameter Estimation

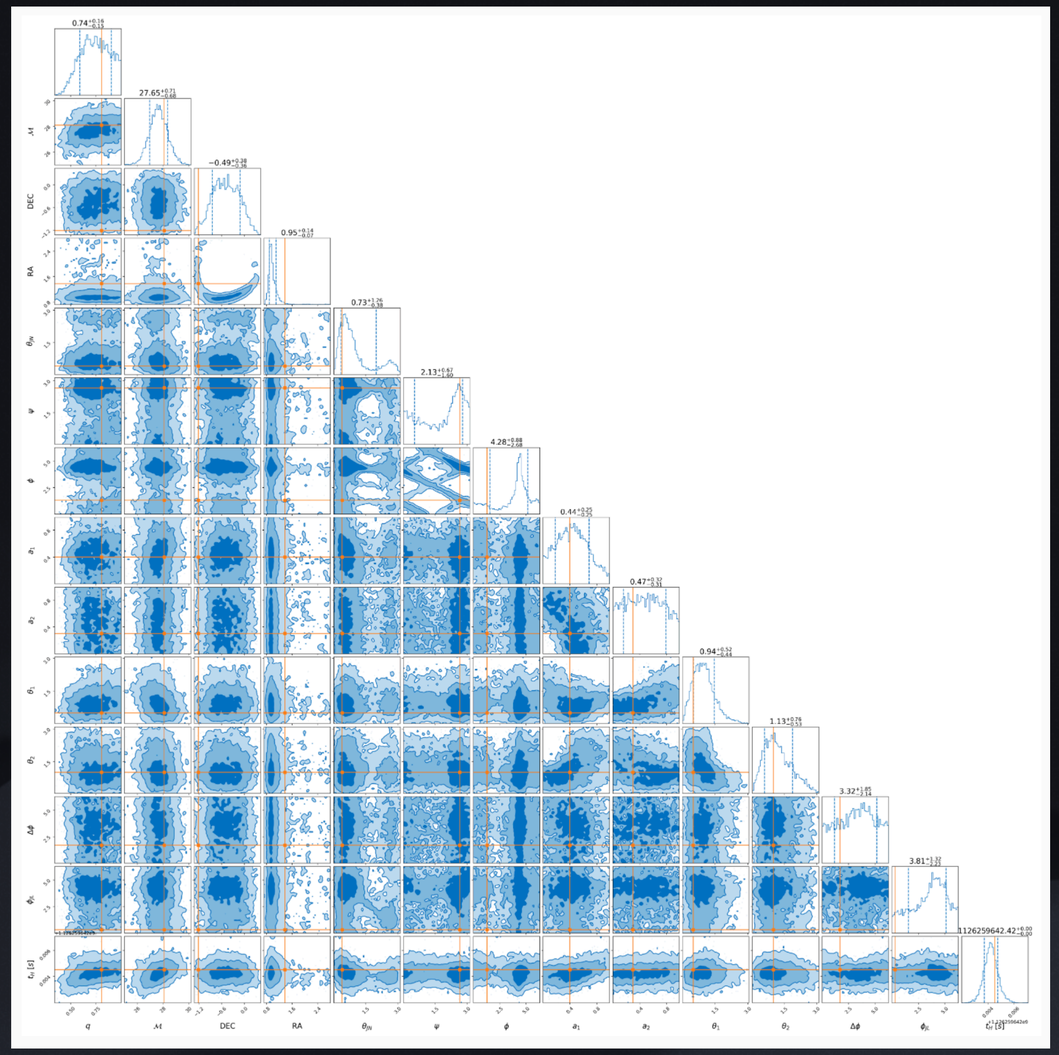
- We also want to know the parameters of the GW sources θ in the strain data d, given the waveform model M and some background knowledge I
- Bayes theorem:

Likelihood Prior

$$P(\theta \mid d, M, I) = \frac{P(d \mid \theta, M, I)\, P(\theta \mid M, I)}{P(d \mid M, I)}$$

Posterior Evidence

- We use the Monte Carlo method to obtain the posterior
- Bilby: a Python package for GW parameter inference



Thanks for Your Attention!

- To learn more about GWs and data analysis:
  - 2026 GW Open Data Workshop @ Toulouse, France
  - Aprl. 20-22, 2026; 2-5 PM CEST (8-11 PM Taiwan time)
  - Online lectures and tutorials
  - Free enrollment on https://gw-odw.thinkific.com/courses/odw2026