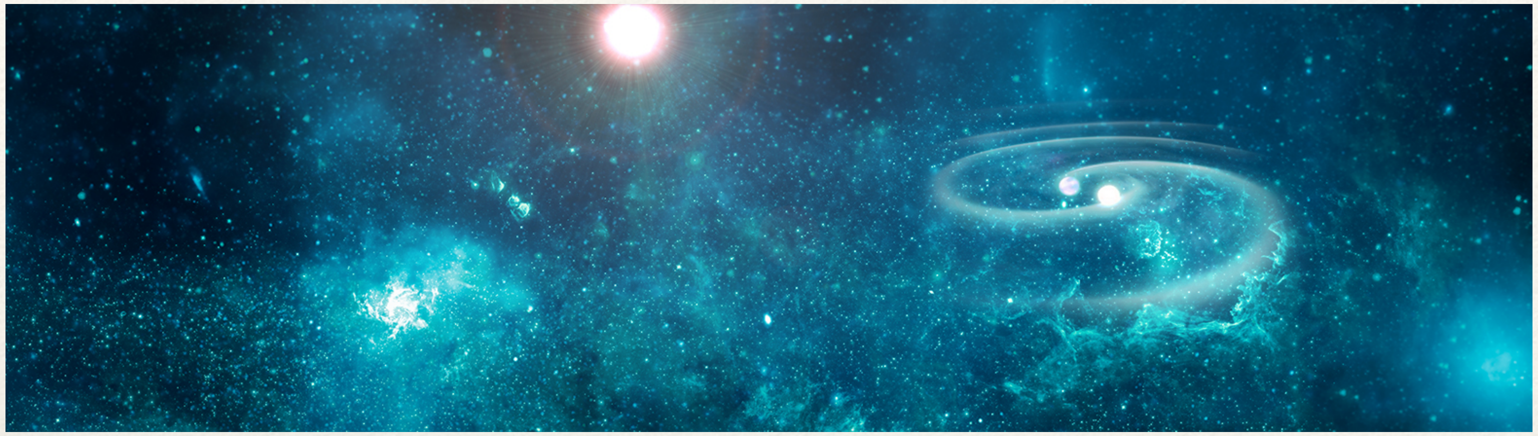
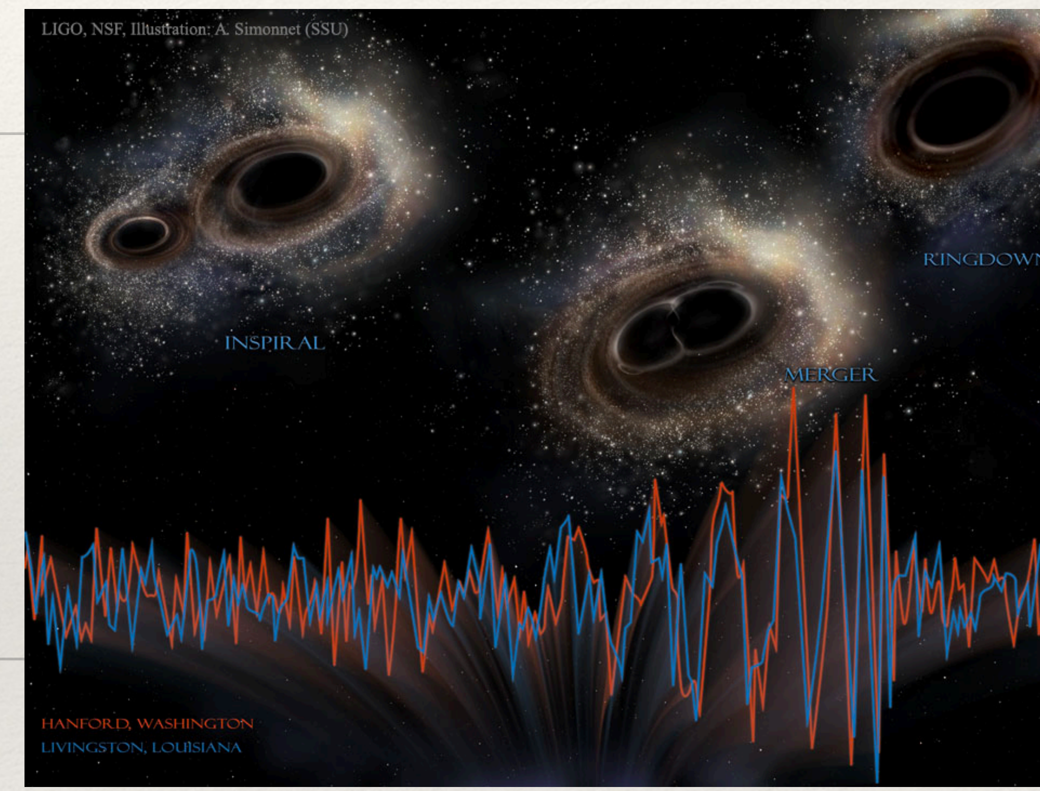
Stanislav (Stas) Babak.

AstroParticule et Cosmologie, CNRS (Paris)



LISA Data Analysis



"Wanderer, your footsteps are the road, and nothing more; wanderer, there is no road, the road is made by walking. By walking one makes the road, and upon glancing behind one sees the path that never will be trod again. Wanderer, there is no road — Only wakes upon the sea." [Antonio Machado]

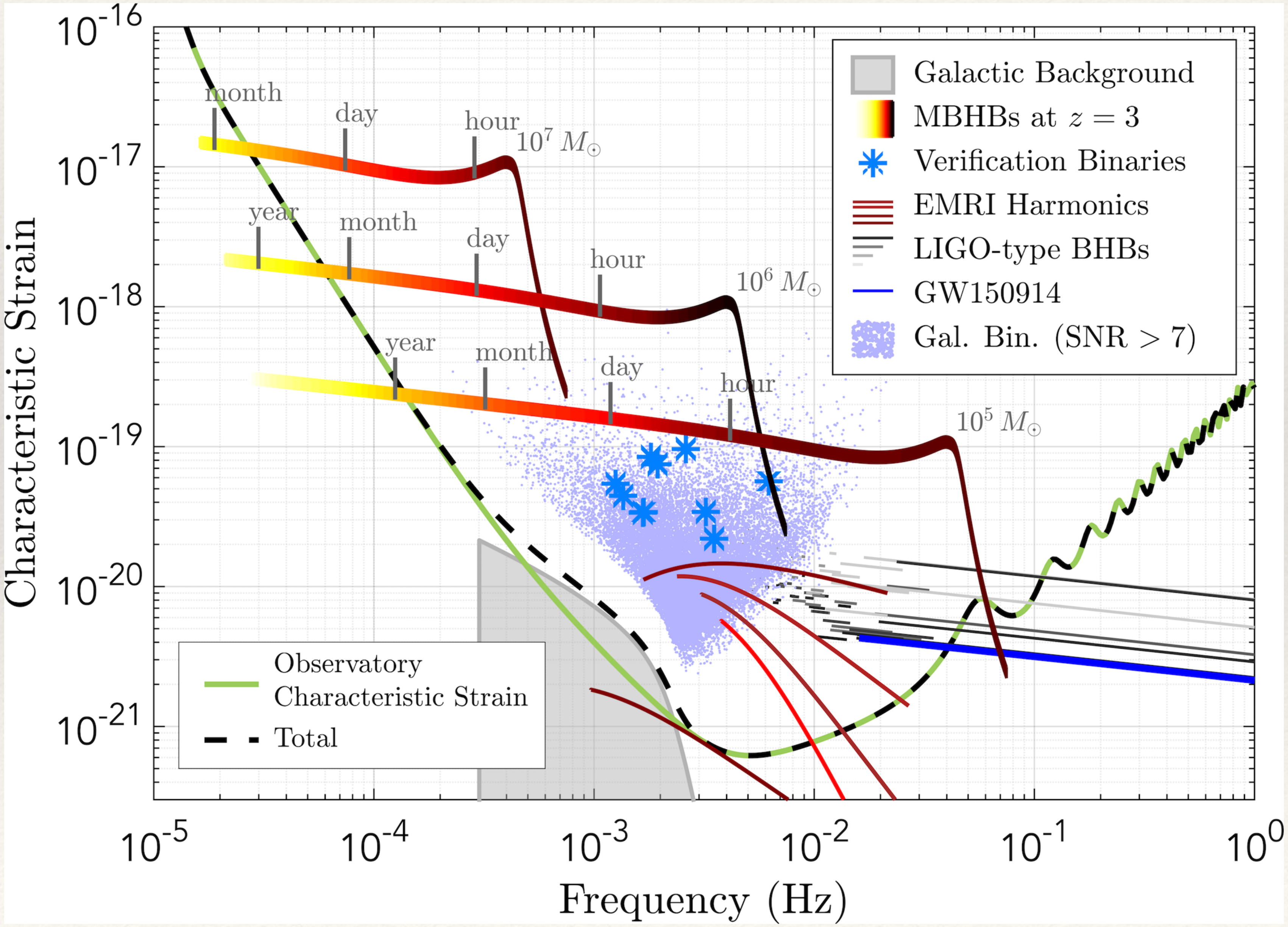
Covid, 7-8 January 2021

Outline

- Gravitational Wave sources in the LISA's band
- Gravitational Wave data analysis
- LISA mock data challenge

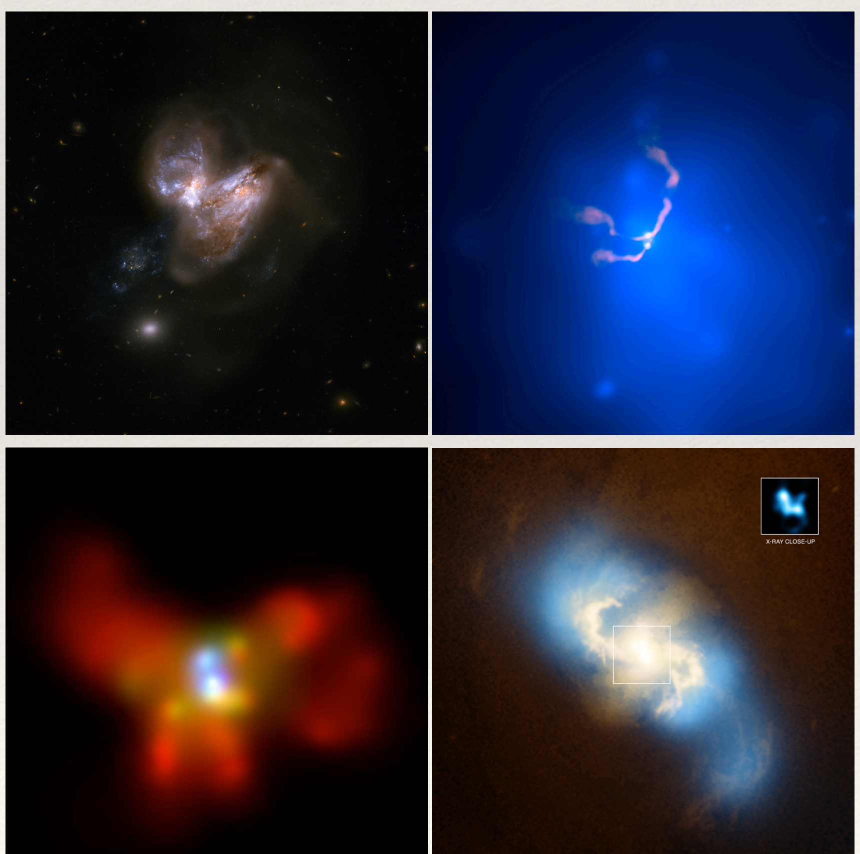
GW sources in LISA band

- GW signals in LISA are strong and long-lived.
- LISA data will contain thousands of GW signals simultaneously: need to separate and characterize them
- Non-stationary noise



LISA sources

- We believe that all galactic nuclei host Massive Black Holes: Milky Way has 4 mln. solar mass BH
- Galaxies merge: we can form Massive Black Hole Binary (MBHB) system
- We need stars and gas to bring MBHs close together for GW to be efficient (binary is merging within Hubble time)

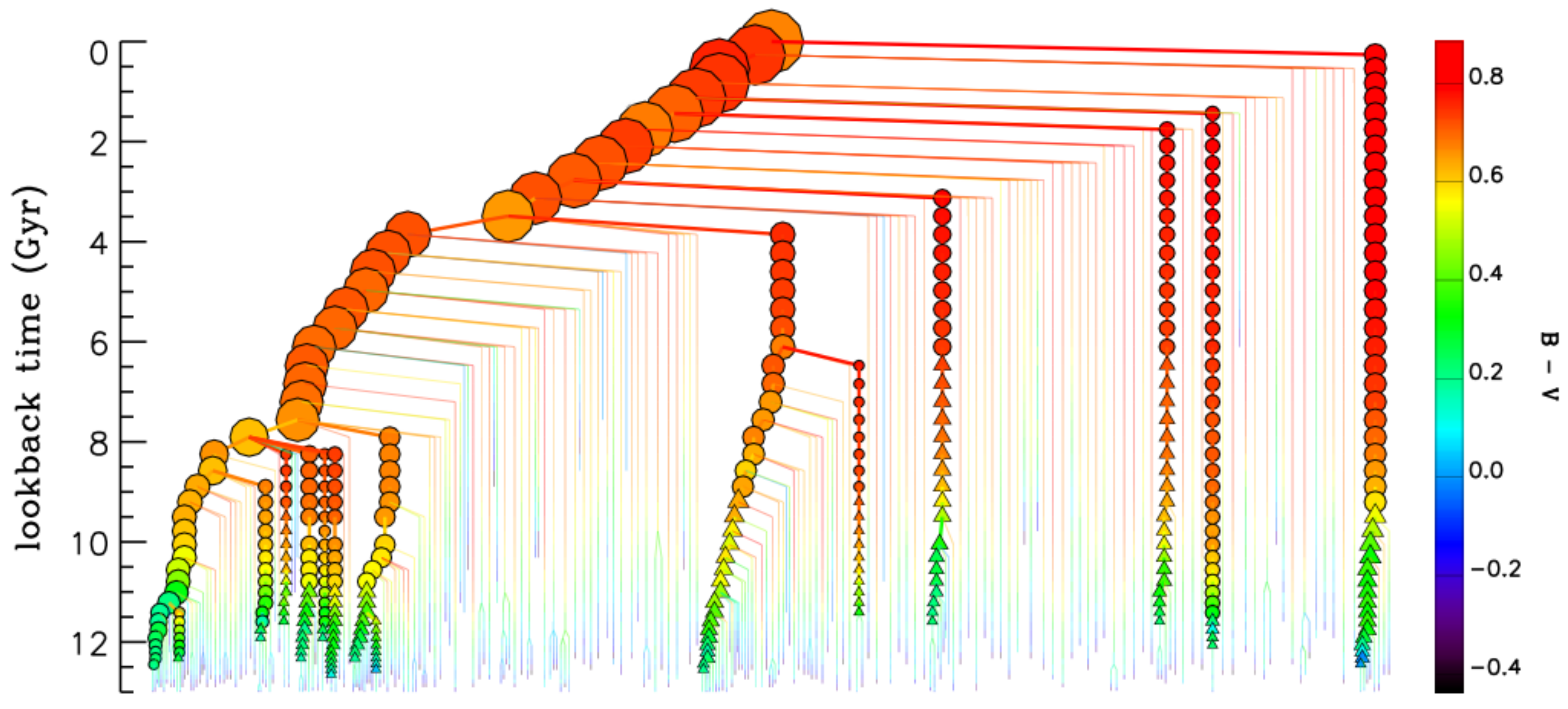


[Credits: Hassinger+, VLA, Chandra , NASA]

[Image: Hubble telescope]

LISA sources: MBHB

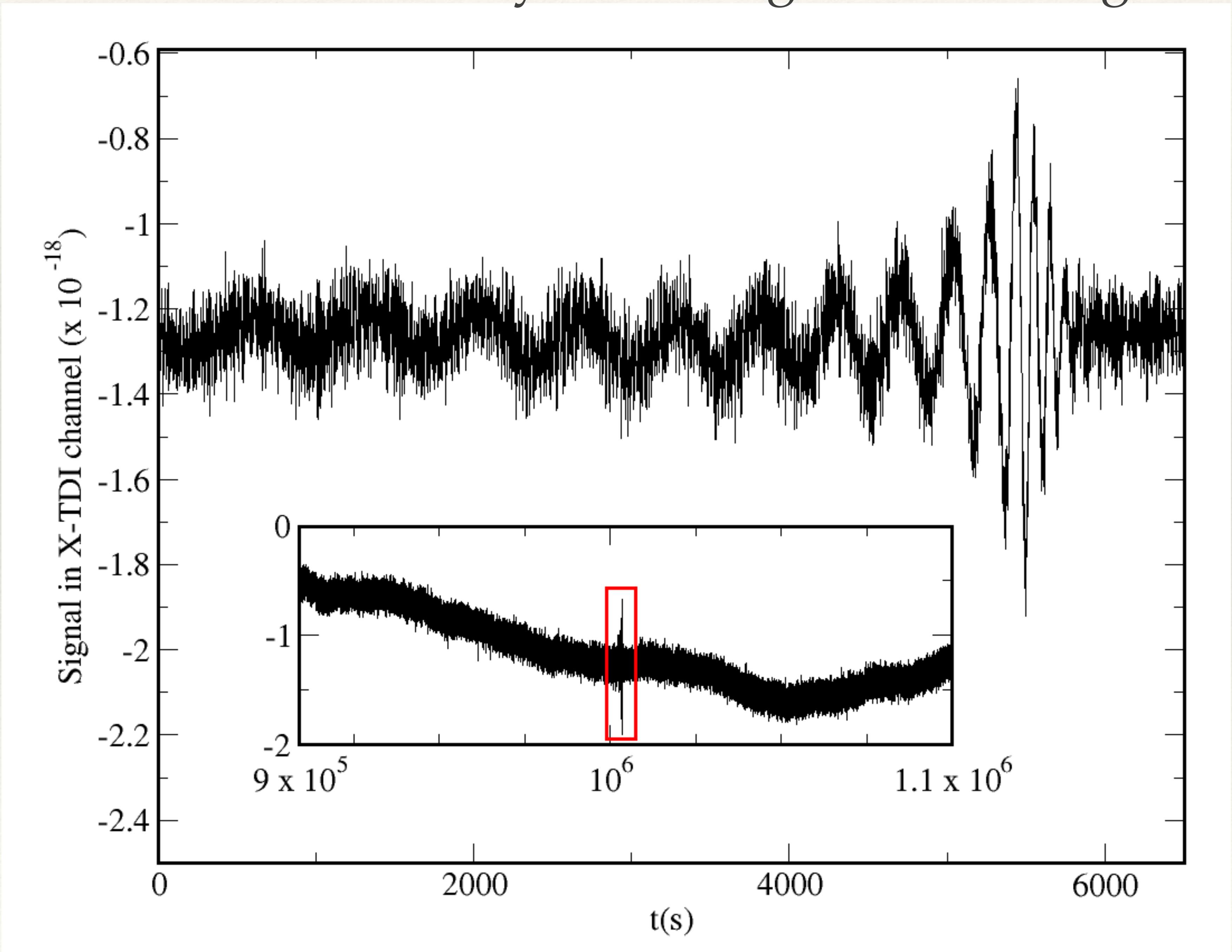
MBHs are formed from the initial BH seed. Those seeds could be “light” remnant of the first generation of stars or “heavy” from the direct collapse of a giant gas cloud. BHs accumulated the mas through gas accretion and merging.



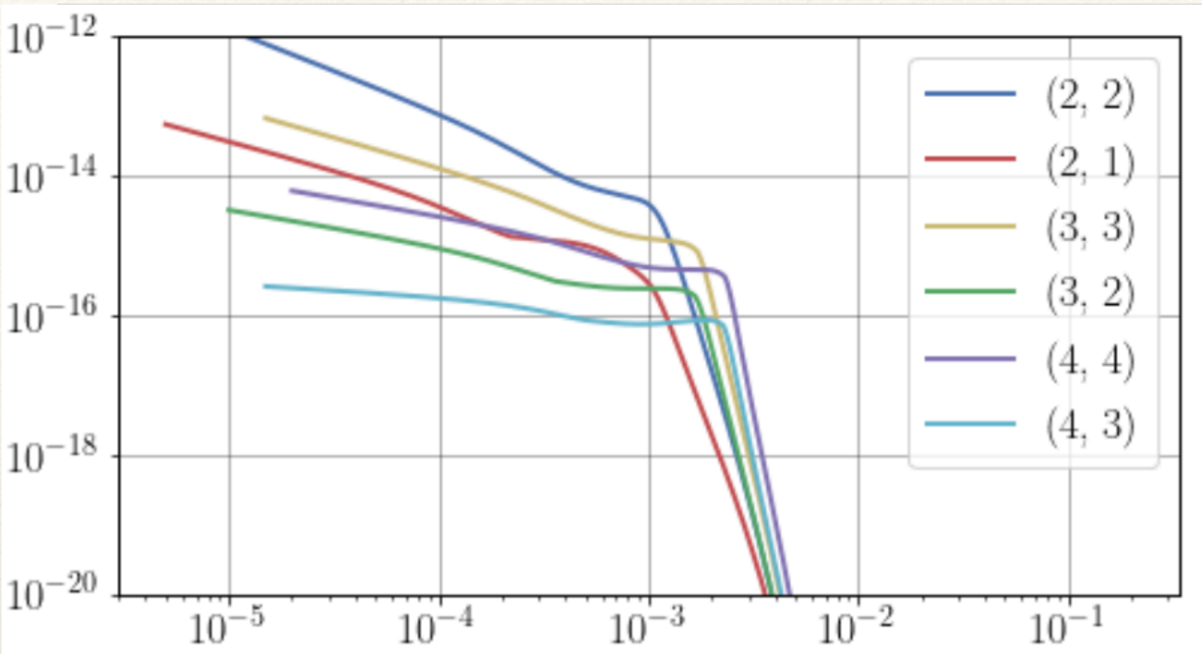
[Credits: Gabriella De Lucia]

LISA: GW signal from MBHB

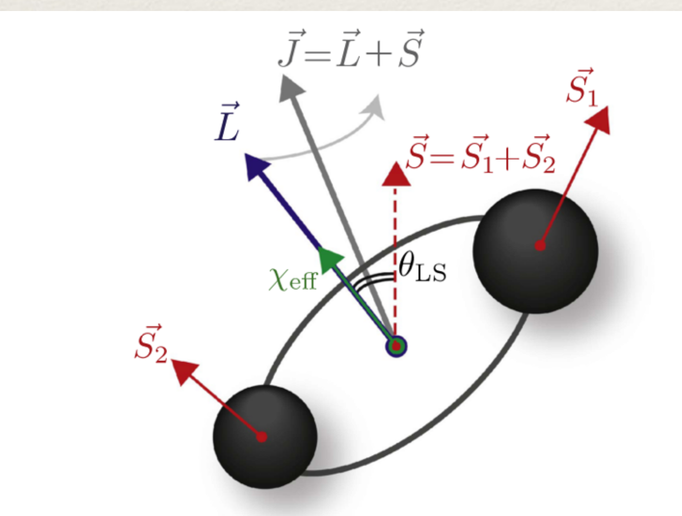
The signal from MBHB is similar to what we have observed in LIGO (scaled up in the amplitude and streched in time). GW signal from MBHB is expected to be the strongest signal (seen by eye in the simulated data). Imposes stringent demands on the accuracy of GW signal modelling



LISA: GW signal from MBHB

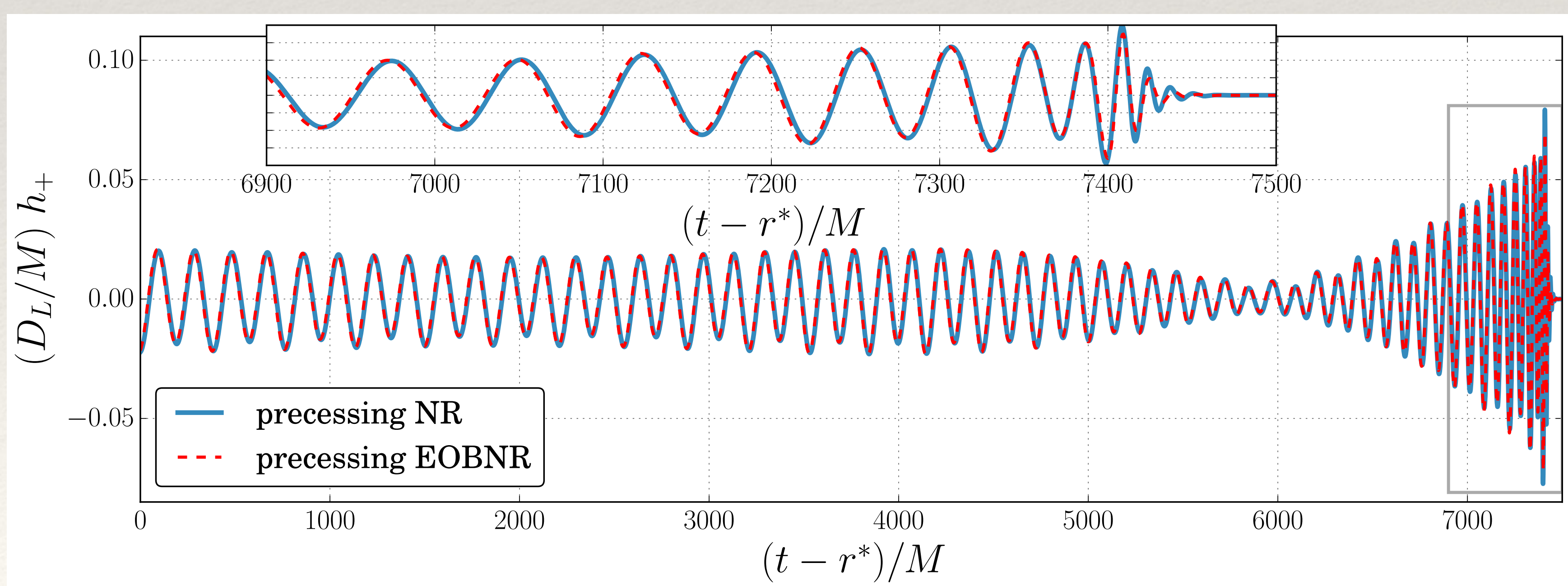


- Signal is superposition of harmonics:
  - dominant is at 2 x orbital frequency
  - subdominant: higher order modes beyond quedrupole approximation



Credit: Carl Rodriguez

- If spins are not paralle to the orbital angular momentum:
  - orbital precession around total momentum of system



[Babak+ (2016)]

EMRIs (extreme mass ratio inspirals)

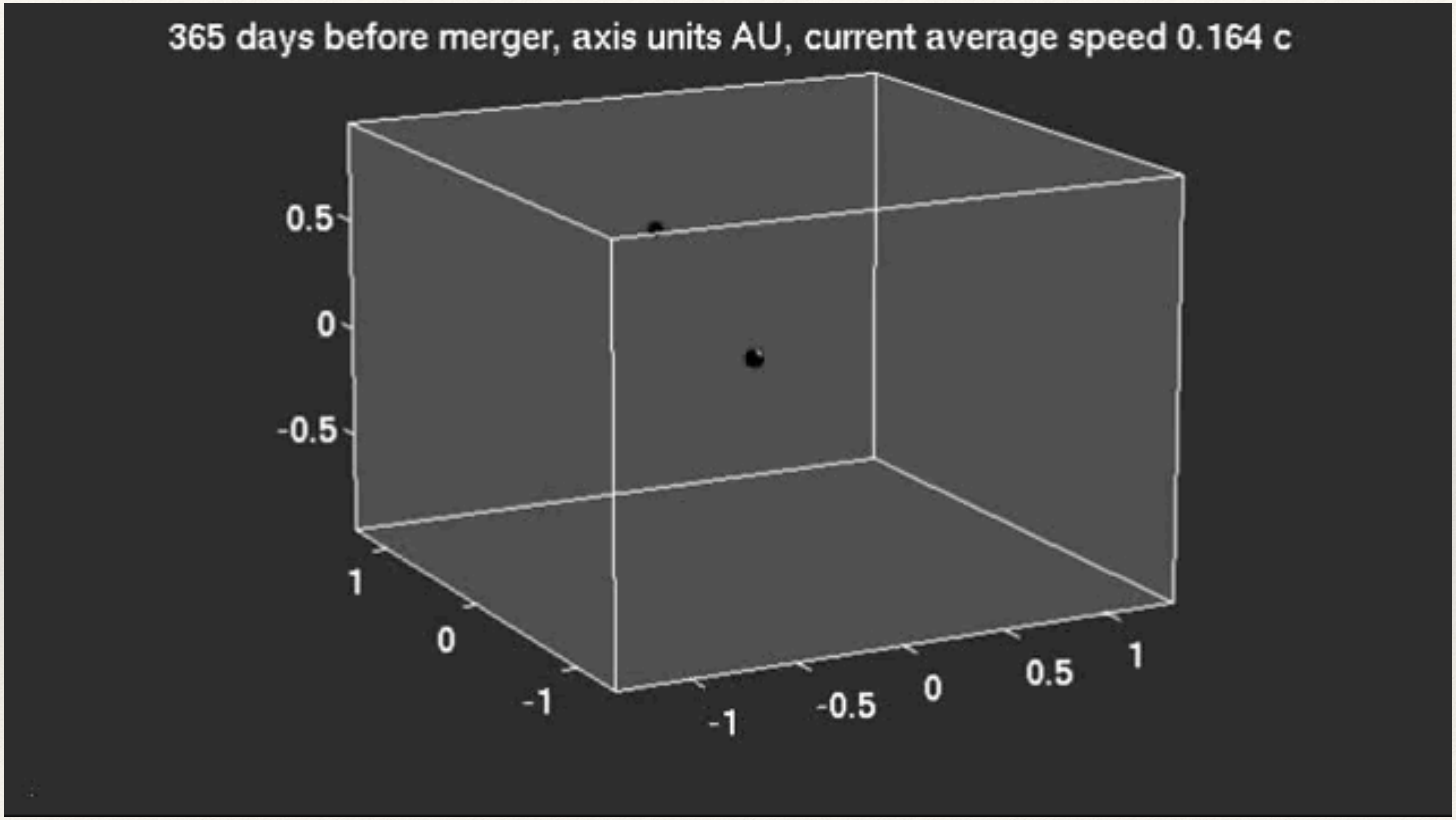
- Massive BHs in galactic nuclei surrounded by stars and gas with quite high density

- MBH could capture a compact object (BH, NS, WD) which is thrown on a very eccentric orbit (due to N-body interaction). The orbit shrinks and circularizes due to grav. radiation.

- EMRI: binary system with extreme mass ratio of component 10^{-7} - 10^{-5}

- Compact object revolves 10^6 orbits in the proximity of MBH before the plunge.

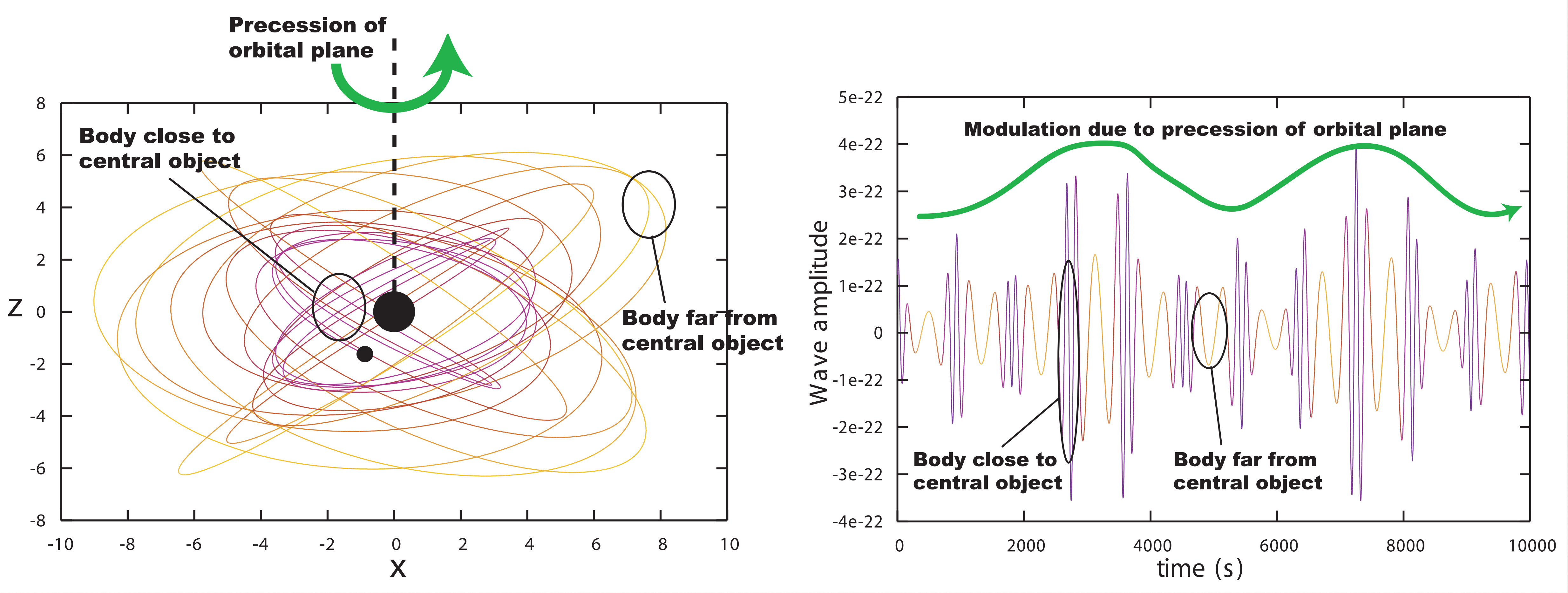
EMRI



[Credits: S Draco, CalTech]

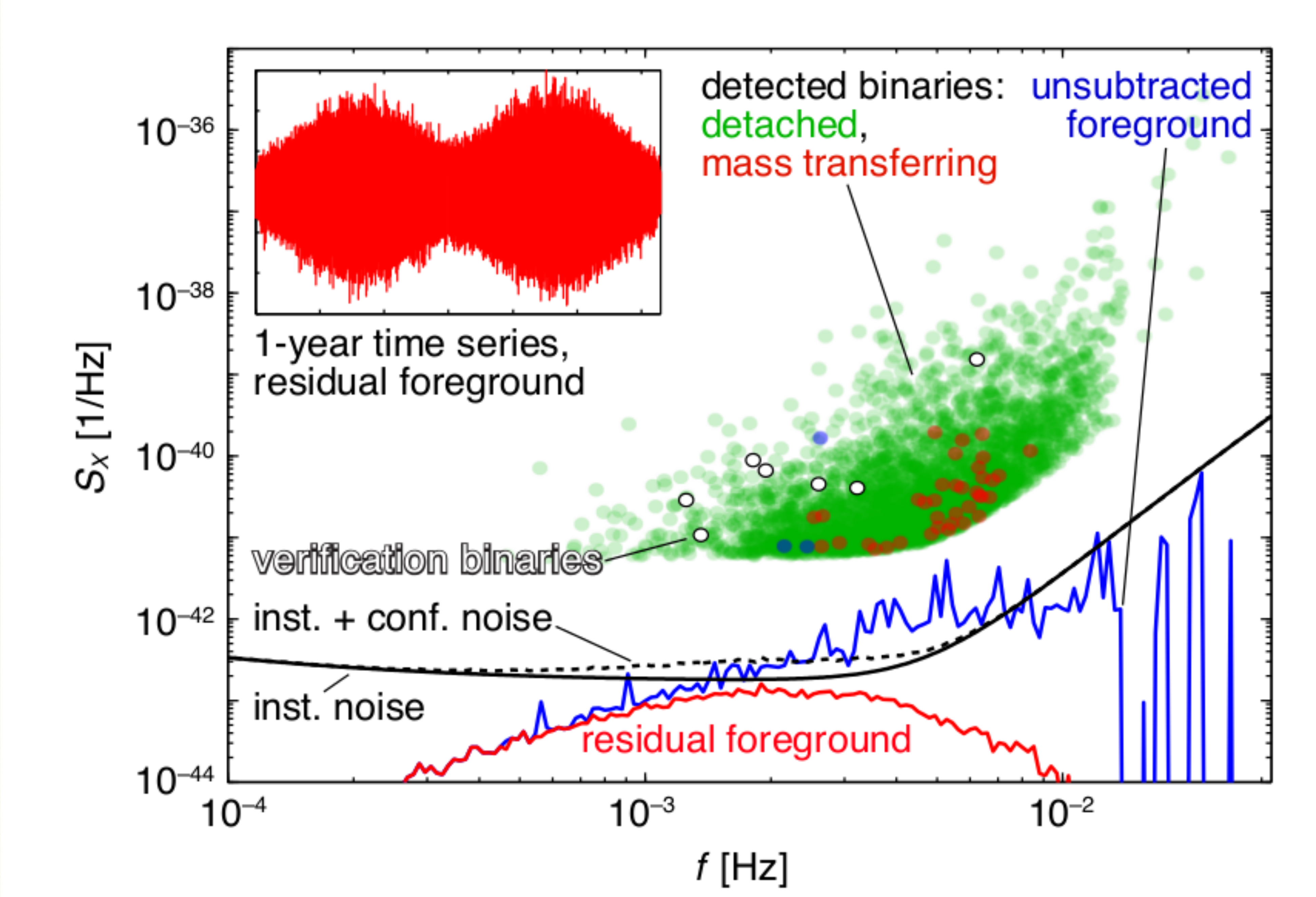
EMRI

- Orbital motion: (almost) elliptical with a strong relativistic precession + orbital precession due to spin-orbital coupling
- Signal is very rich in structure (hard to detect but gives a lot of information)
- Ultra-precise parameter determination (if detected). Can map spacetime of a heavy object: holiodesy



Galactic white dwarf binaries

- We expect to have 10^7 WD binaries all emitting GWs in the LISA band, only 10^4 can be resolved individually, other form stochastic GW signal (foreground)
- GW signal is almost monochromatic
- Verification binaries: known from current e/m observations (+GAIA,+ LSST)



Expected event rate in LISA

- MBHB : high uncertainties in the event rate - from few to few hundreds per year

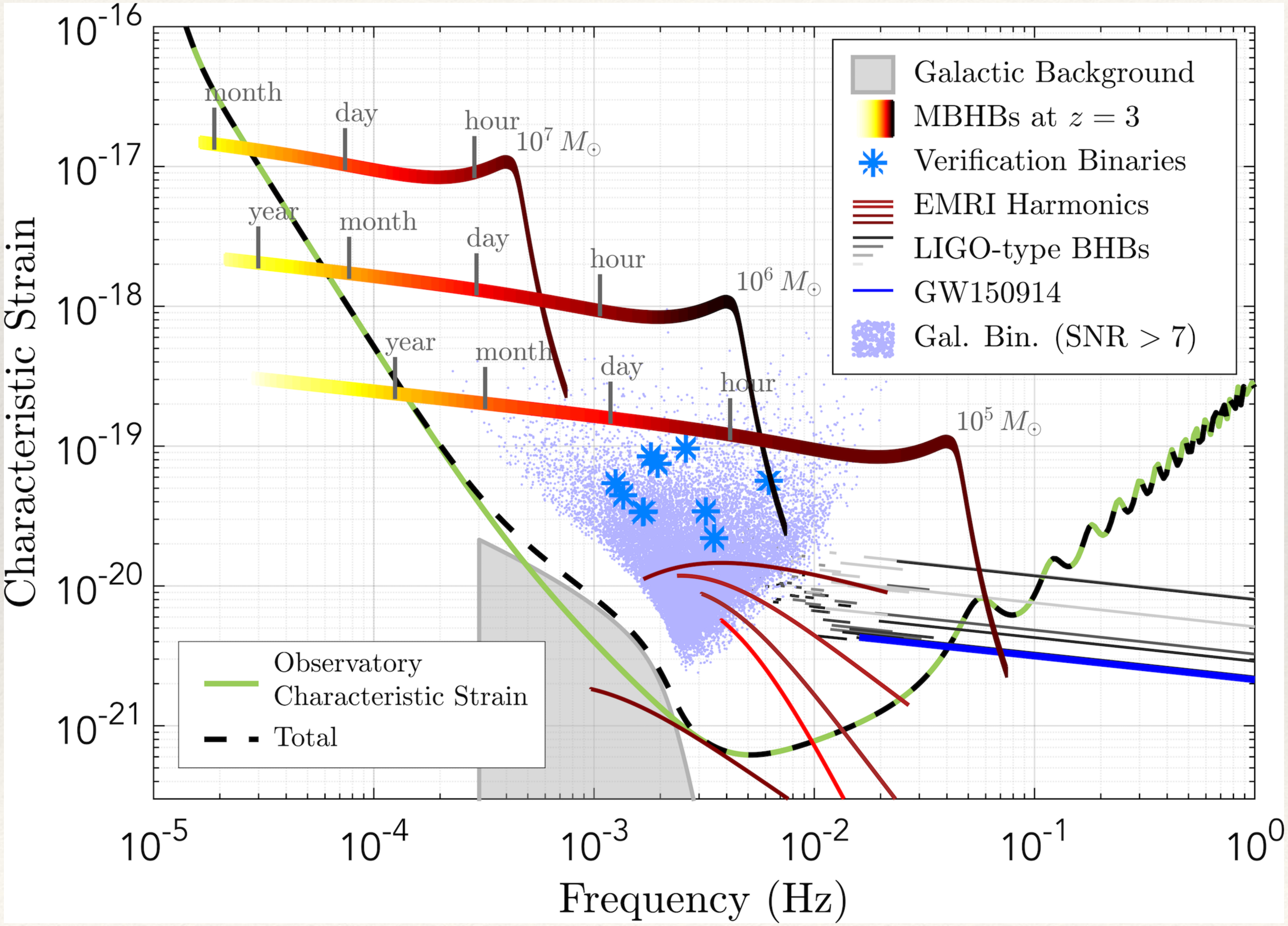
- EMRIs: even more uncertain - from few to few thousands of detectable GW signals per year.

- GW signal from solar mass BBH (LIGO/VIRGO sources). We expect to observe about 10 sources: GW signal first observed in LISA and then 5-10 years later with the ground based detectors.

- Possible detection of the stochastic GW signal from energetic processes in the early Universe.

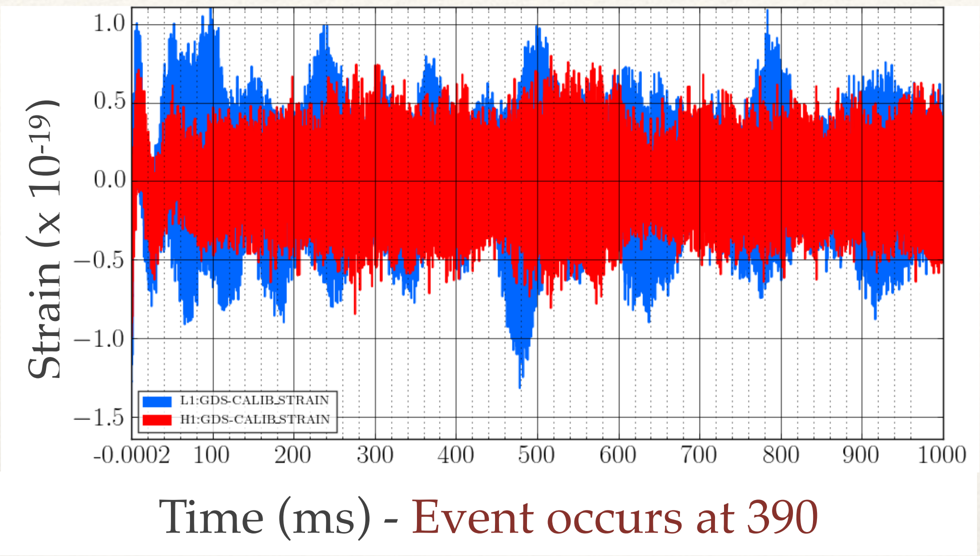
LISA data

LISA data analysis is quite a complex task. We organize the LISA data challenge: https://lisa-ldc.lal.in2p3.fr/home . We simulate LISA data (noise and GW signals) and anyone can download the data and analyse it.

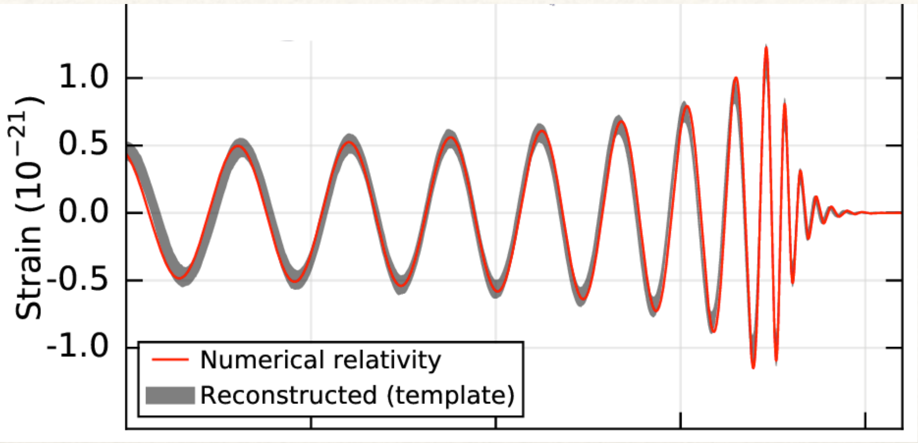


Data analsys: Matched filtering

GW150914 Raw data



GW signal from merging BHs (we search for)



Matched filtering: is used when we are searching for a signal of known form in the noisy data. The basis: we correlate the data with expected signal and search for a maximum of correlation.

wavform / template we search for.

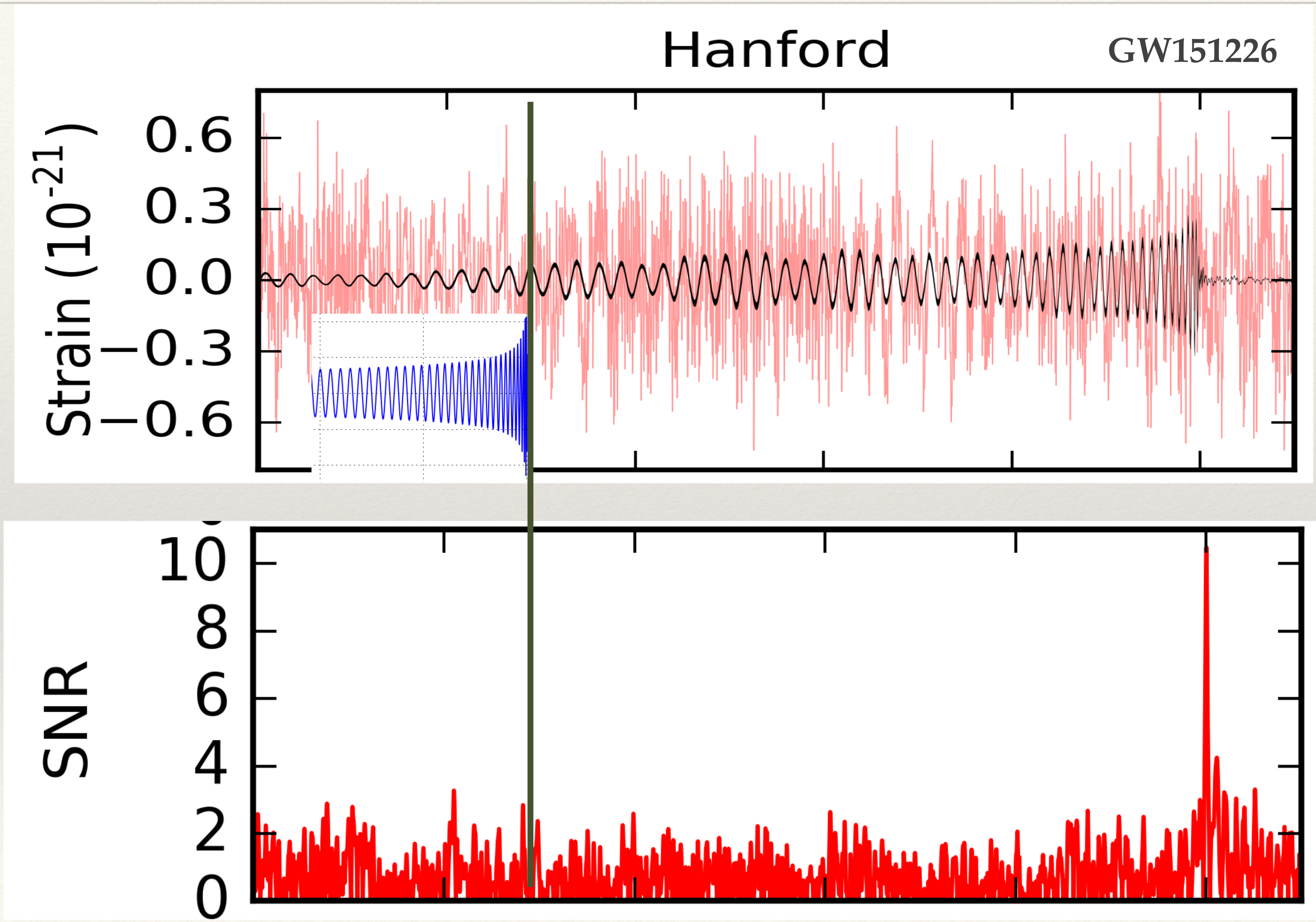
$$\rho \sim 4\Re \int_0^\infty \frac{\tilde{d}(f)\tilde{h}^*(f)}{S(f)} df$$

data

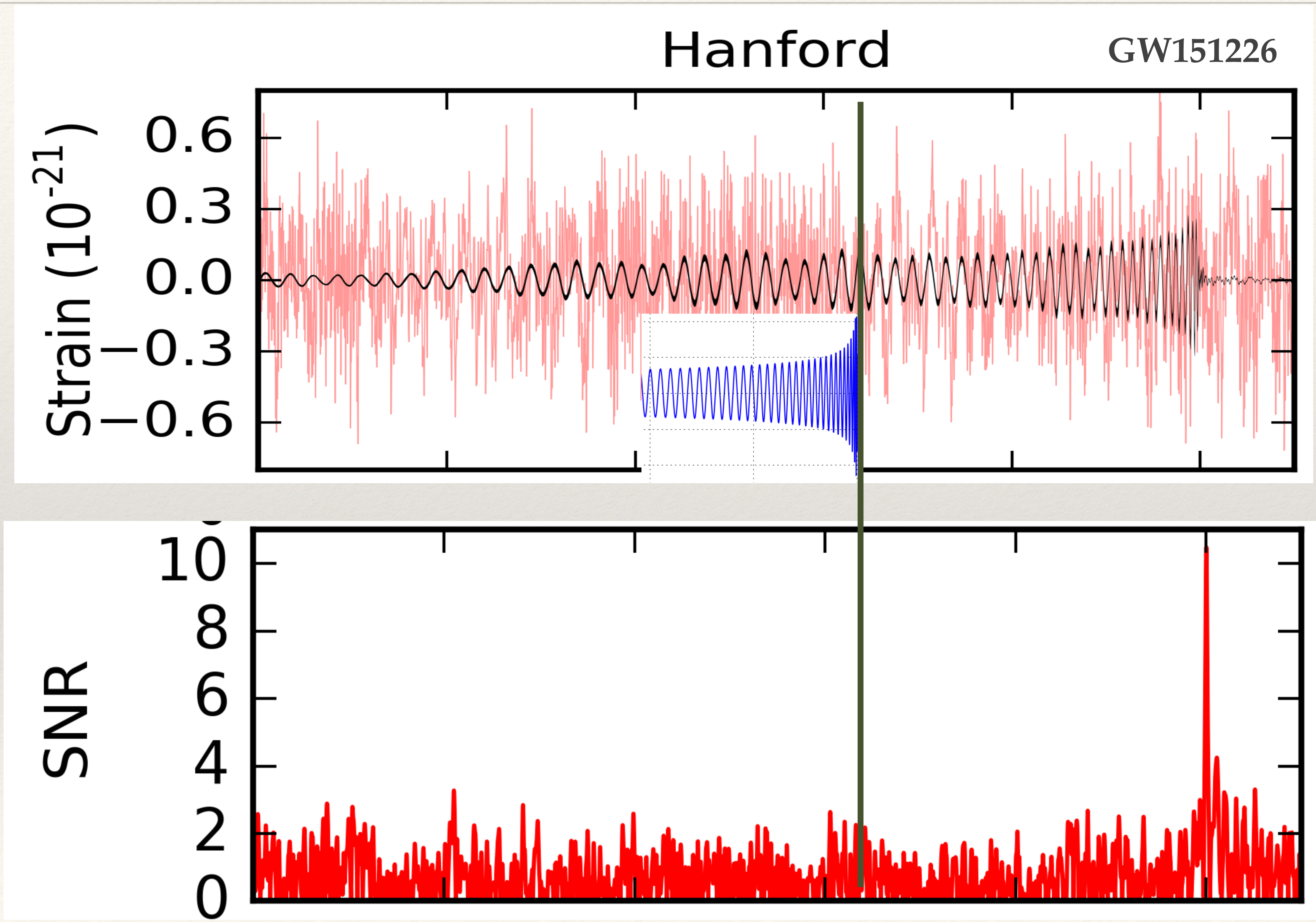
noise power spectral density

Correlation in frequency domain, weighted by detector sensitivity

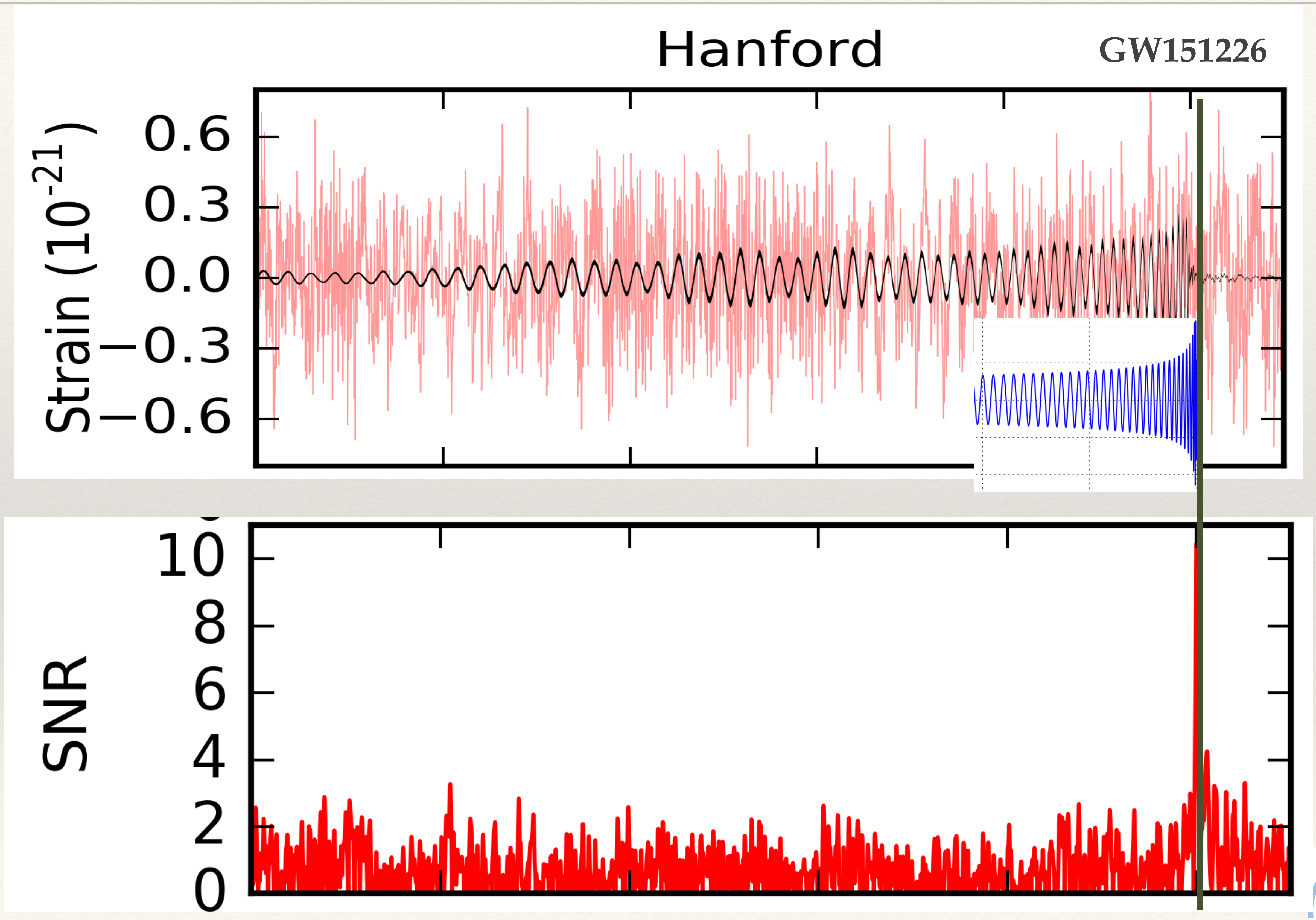
Matched filtering



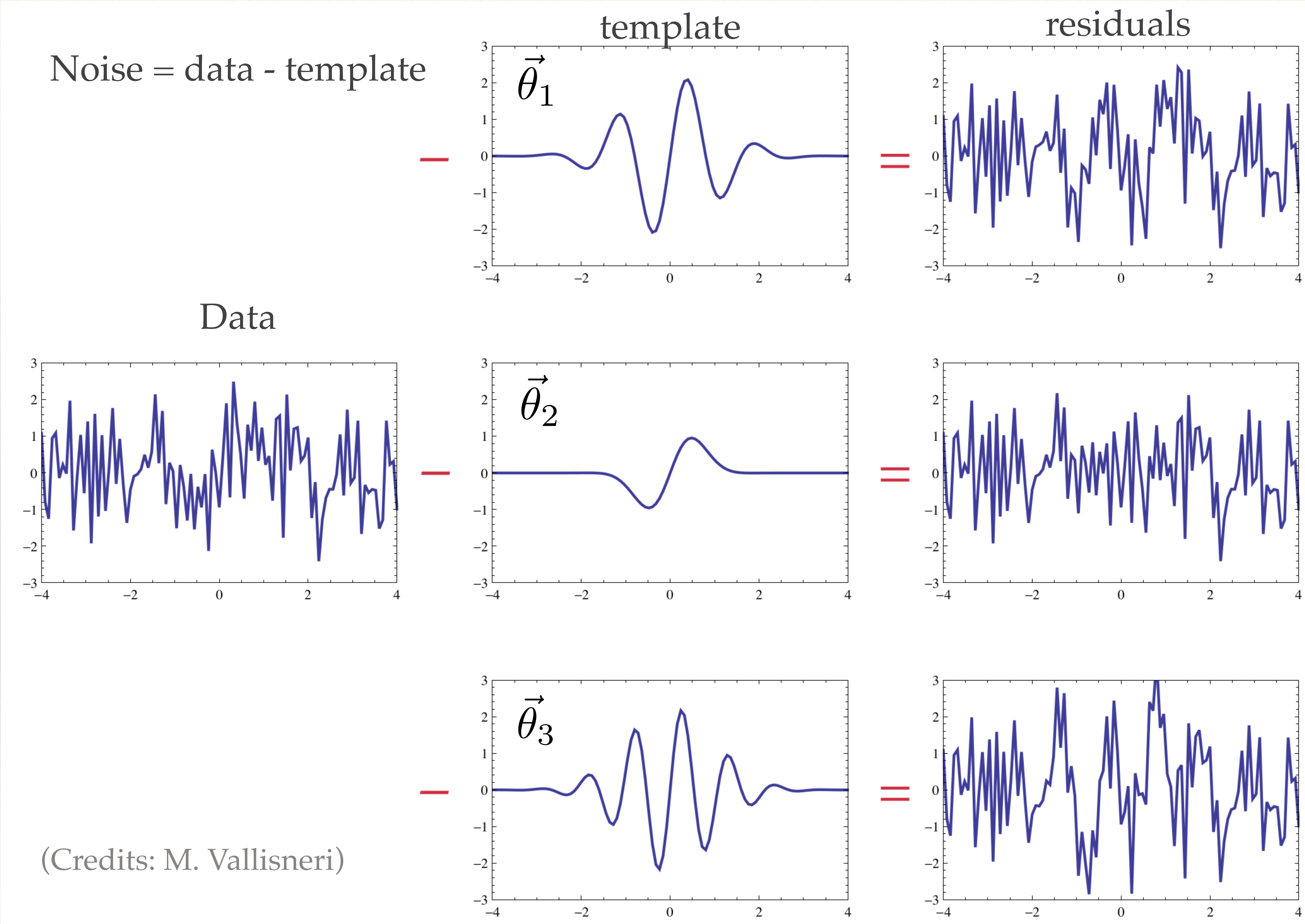
Matched filtering



Matched filtering

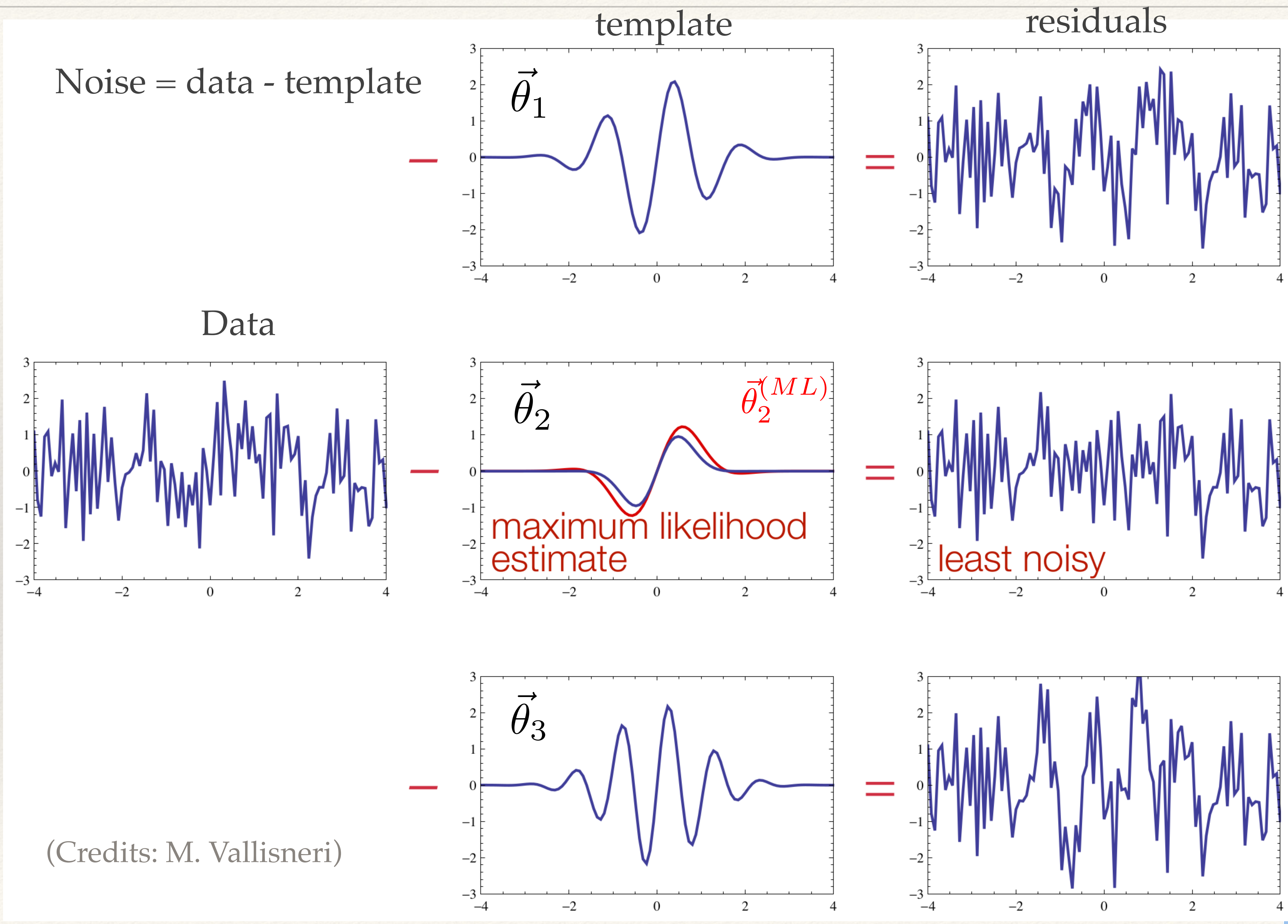


Matched filtering and parameter estimation



(Credits: M. Vallisneri)

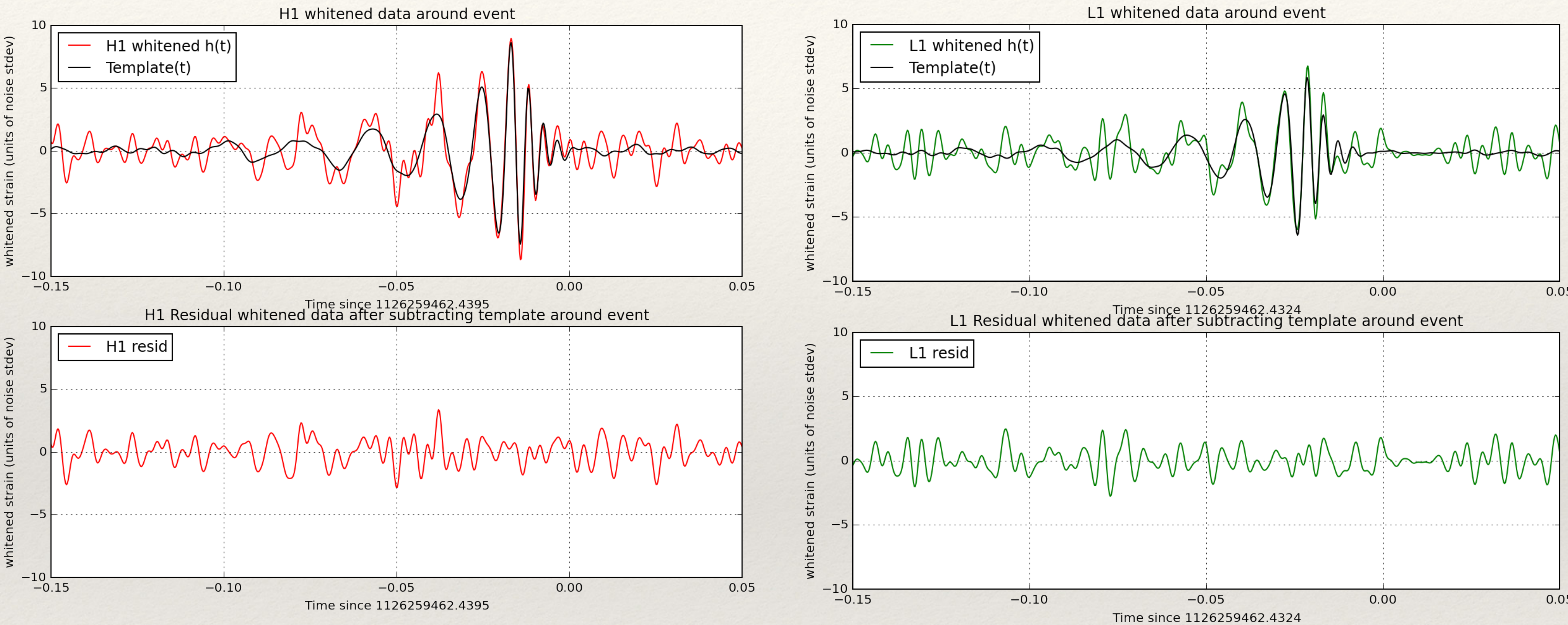
Matched filtering and parameter estimation



(Credits: M. Vallisneri)

Matched filtering for GW150914

H1 **L1**



[LOSC: https://losc.ligo.org/tutorials/]

Likelihood

Let us assume that the data contains the signal.: hypothesis (model) H_1

$d(t) = n(t) + s(t, \theta_i)$ signal "s" depends on parameters θ_i

data = noise + signal

If the template matches the GW signal exactly $h(t, \theta_i) = s(t, \theta_i) \longrightarrow d(t) - h(t, \theta_i) = n(t)$

$$p(d(t)|H_1, \bar{s}(t, \lambda)) = p(d(t) - s(t, \vec{\lambda})) = p_n$$

- Assume that the noise is Gaussian (but not necessarily white): non white noise has different variance at different frequencies. The the likelihood can be written as

Likelihood: $p(d|H_1, \theta_i) \propto e^{-\frac{1}{2}(d-h(\theta_i)|d-h(\theta_i))}$

The inner product: matched filtering $$(a|b) \equiv 4\Re \int_0^\infty \frac{\tilde{a}(f)\tilde{b}^*(f)}{S_n(f)}\, df$$

- We search for parameters which maximize the likelihood: making the residuals most noise-like — maximum likelihood estimators for parameters $\hat{\theta}_i$

Likelihood

- For a given noise realization the maximum likelihood estimators. $\hat{\theta} \neq \theta_{true}$

How close the estimated parameters to the true depends on the noise realization and on the strength of the signal: stronger the signal (high signal-to-noise ratio, SNR) closer estimators to the true values — less influence of the noise. Unbiased estimator: equal to true if averaged over the noise realizations. $< \hat{\theta} >= \theta_{true}$

- If $s \neq h(\theta_{true})$ — lack of accuracy in the signal modelling: systematic bias in parameter estimation.

$$min_{\theta_i}(s - h(\theta_i)|s - h(\theta_i)) \rightarrow \tilde{\theta}_i \qquad |\theta_{true} - \tilde{\theta}_i| = \delta\theta_i \quad \text{bias}$$

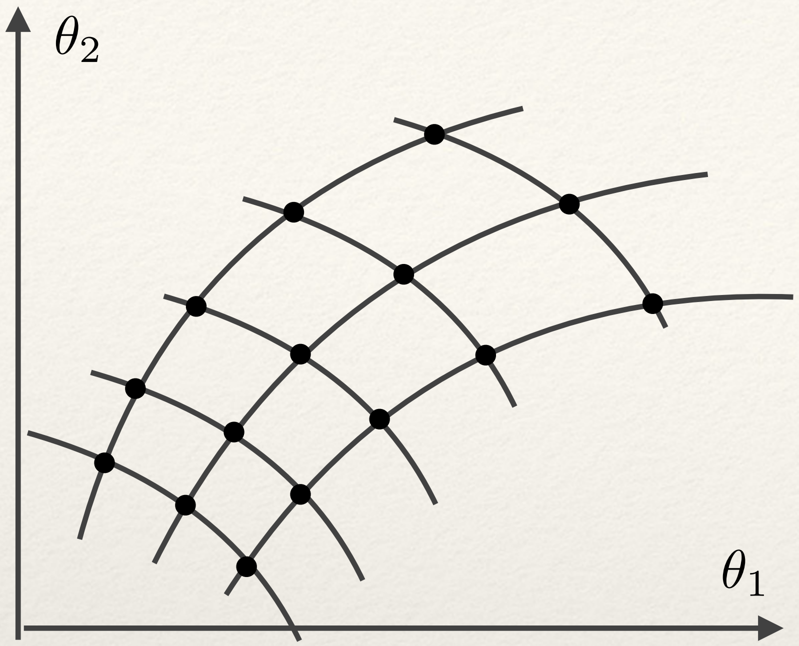
If there is a bias, we still can detect the GW signal on expense of making error in parameters charactrizing the system (binary): *effectualness*

Faithfulness $\dfrac{(s|h(\theta_{true})}{\sqrt{(s|s)(h(\theta_{true})|h(\theta_{true}))}}$ *Overlap* $(\hat{s}|\hat{h}), \quad \hat{h} = \dfrac{h}{\sqrt{(h|h)}}$

$\hat{h}$ → normalized

Overlap varies [-1, 1]: 1 is a pefect match, related to the loss in SNR

Likelihood maximization

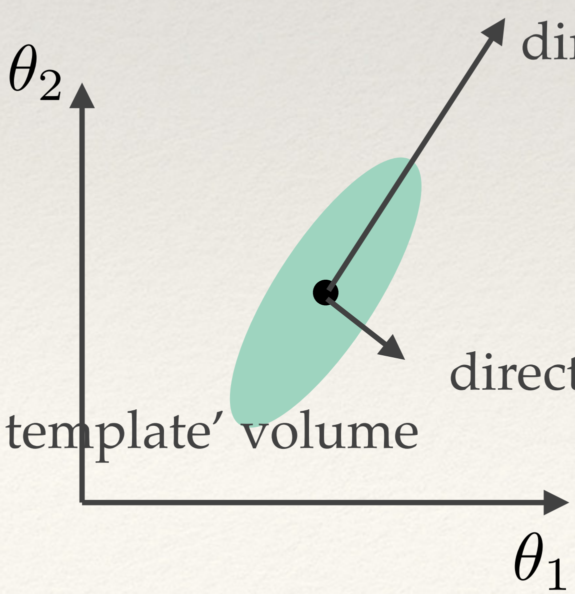


- We want to cover the parameter space (N-dim) by grid of points at equal distance from each other.
- Grid: not too coarse, not too fine
- The distance is determined **not** by a coordinate distance but by “proper” distance — correlation between nearby templates: introduce interval and metric

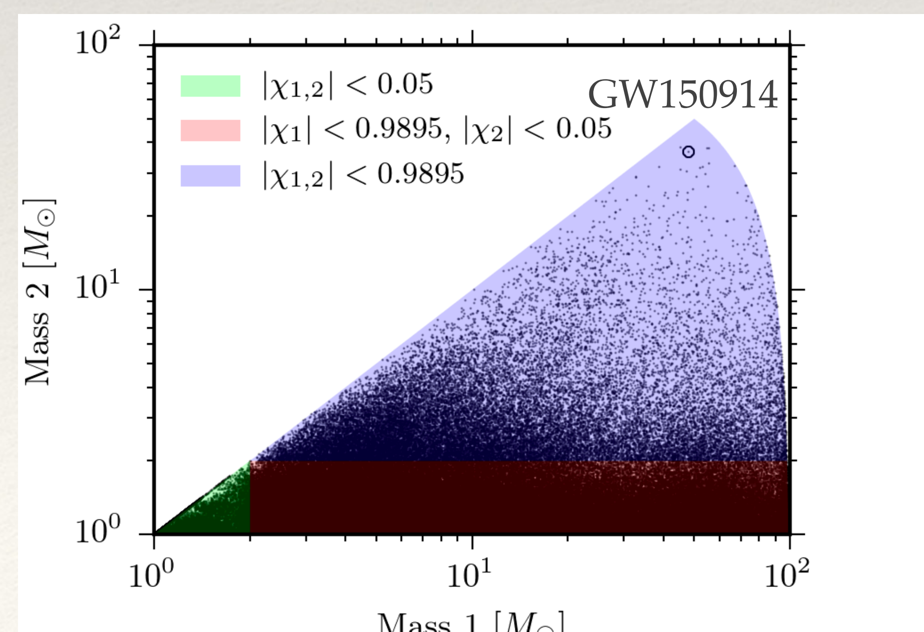
$$ds^2 = |\hat{h}(\theta_i + \delta\theta_i) - \hat{h}(\theta_i)| \approx (\hat{h}(\theta_i + \delta\theta_i) - \hat{h}(\theta_i)|\hat{h}(\theta_i + \delta\theta_i) - \hat{h}(\theta_i)) \approx \left(\frac{\partial \hat{h}}{\partial \theta_i}|\frac{\partial \hat{h}}{\partial \theta_j}\right)\delta\theta_i\delta\theta_j$$

metric on the parameter maniold

Consider 2-D parameter space and fix ds = 0.01



the size and orientation of the ellipse is a function of a central pont



Bayesian approach

- Expensive computationally: often used when the signal is detected using the grid-based method (or something else) : H_1 is true. Allows to test several models (different signal' models, non-GR theories).
- We have to assign the ***prior*** probability to our models and parameters of each model. We treat parameters describing a signal as random variables and trying to estimate probability distribution function(s) for each parameter based on the observed data (***posterior***)
- Consider several models M_i each parametrized by set of parameters $\vec{\theta}_i$

Bayes' theorem $$P(M_i|d) = \frac{P(d|M_i)\pi(M_i)}{p(d)}$$

For a given model M_i

likelihood prior

$$P(\vec{\theta}_i|M_i, d) = \frac{P(d|\vec{\theta}_i, M_i)\,\pi(\vec{\theta}_i)}{p(d|M_i)}$$

posterior

Evidence of model M_i

Bayesian approach

$$p(d|M_i) = \int d\vec{\theta}_i \; p(d|\vec{\theta}_i, M_i)\pi(\vec{\theta}_i)$$

— Evidence of model M_i: important for the model selection

$$P(M_i|d) = \left[\int d\vec{\theta}_i \; p(d|\vec{\theta}_i, M_i)\pi(\vec{\theta}_i)\right] \frac{\pi(M_i)}{P(d)}$$

— probability of the model M_i given a data

- **Odd ratio**:The problem to evaluate the normalization P(d) - requires full set of models which are mutually exclusive. We can evaluate the ratio of probabilities:

$$O_{a,b} = \frac{p(M_a|d)}{p(M_b|d)} = \frac{p(d|M_a)}{p(d|M_b)} \frac{\pi(M_a)}{\pi(M_b)}$$

Bayes factor prior odds

Markov Chain Monte Carlo (MCMC)

- So for a given model M_i we need to evaluate postrior pdf for all parameters and the evidence: posterior pdf tells us about parameters of the GW signal (system) and evidence tells us how good this models fits the observations.

$$p(\vec{\theta}_i|M_i, d) = \frac{p(d|\vec{\theta}_i, M_i)\ \pi(\vec{\theta}_i)}{p(d|M_i)}$$ — posterior pdf

$$p(d|M_i) = \int d\vec{\theta}_i\ p(d|\vec{\theta}_i, M_i)\pi(\vec{\theta}_i)$$ — evidence of model M_i

- ***Markov Chain Monte Carlo (MCMC)*** approach:

We construct Markov chain: stochastic process where the next point in the chain depends only on the previous one. And:

- we want chain to move towards the region of parameter space with high likelihood
- we need to introduce the transitional probability: way to move from one point to another
- we want a transitional probability to satisfy the ballance equation

ballance eqn.: $$P(\vec{\theta}_{(k)})P(\vec{\theta}_{k+1}|\vec{\theta}_{(k)}) = P(\vec{\theta}_{(k+1)})P(\vec{\theta}_k|\vec{\theta}_{(k+1)})$$

distribution we want to sample (posterior) — transitional probability

MCMC

- Consider a particular implementation: Metropolis-Hastings
  - particular way of building transitional probability which satisfies the ballance equation
  - we start with introducing a proposal distribution (arbitrary*) $q(\vec{\theta}_{(k+1)}|\vec{\theta}_{(k)})$
  - then we build the chain by introducing the acceptance probability

$$\alpha(\vec{\theta}_{(k+1)}|\vec{\theta}_{(k)}) = min\left\{1, \frac{p(d|\vec{\theta}_{(k+1)})}{p(d|\vec{\theta}_{(k)})} \frac{q(\vec{\theta}_{(k)}|\vec{\theta}_{(k+1)})}{q(\vec{\theta}_{(k+1)}|\vec{\theta}_{(k)})} \frac{\pi(\vec{\theta}_{(k+1)})}{\pi(\vec{\theta}_{(k)})}\right\}$$

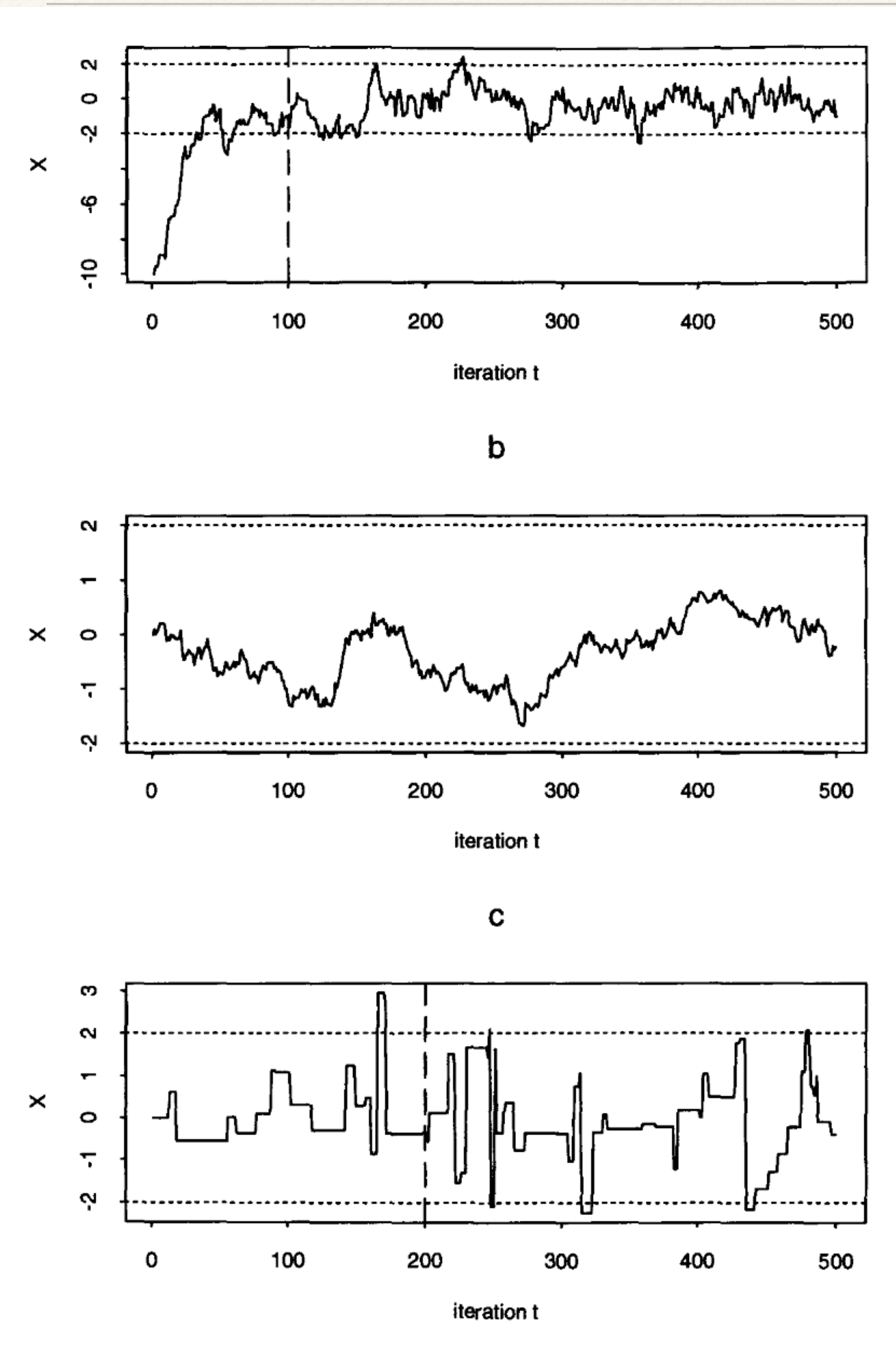
likelihood ratio **ratio of proposals** **ratio of priors**

It is easier to understand if we use symmetric proposal and uniform priors: likelihood ratio

α — probability of accepting new point $\vec{\theta}_{(k+1)}$

The chain moves predominantly in the direction of high likelihood.

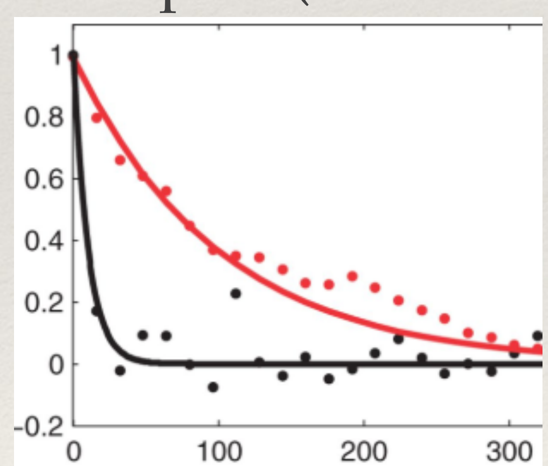
MCMC



- The theorem tells us that the chains will sample the posterior pdf (after some burn-in length) independent of the proposal distribution, BUT

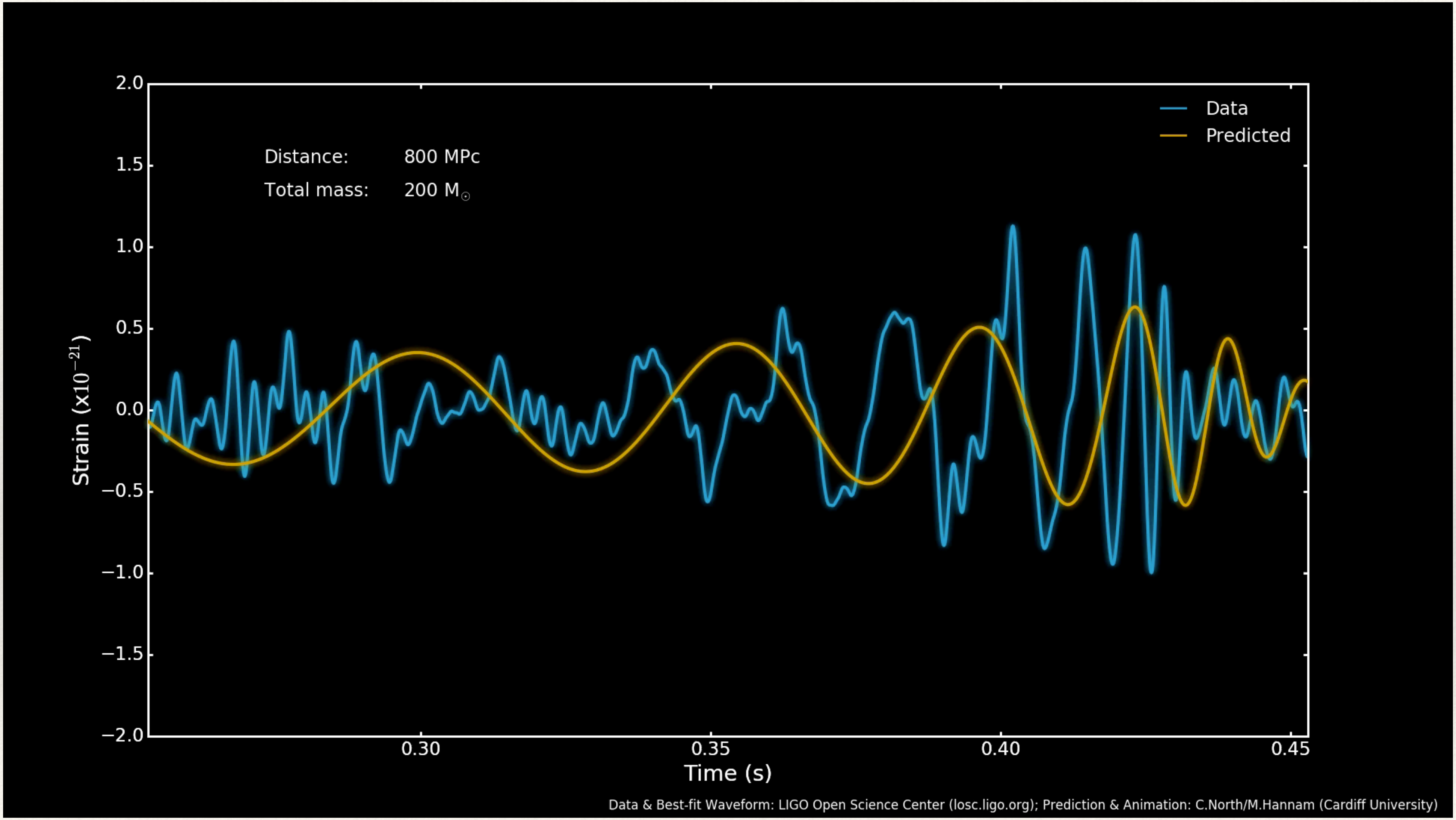
- The efficiency of the sampling strongly depends on the proposal (proposal should resemble the posterior)

- Number of samples vs. number of independent samples (defined by autocorrelation length)



- Multimodal posterior require special treatment! (simulated annealing, parallel tempering)

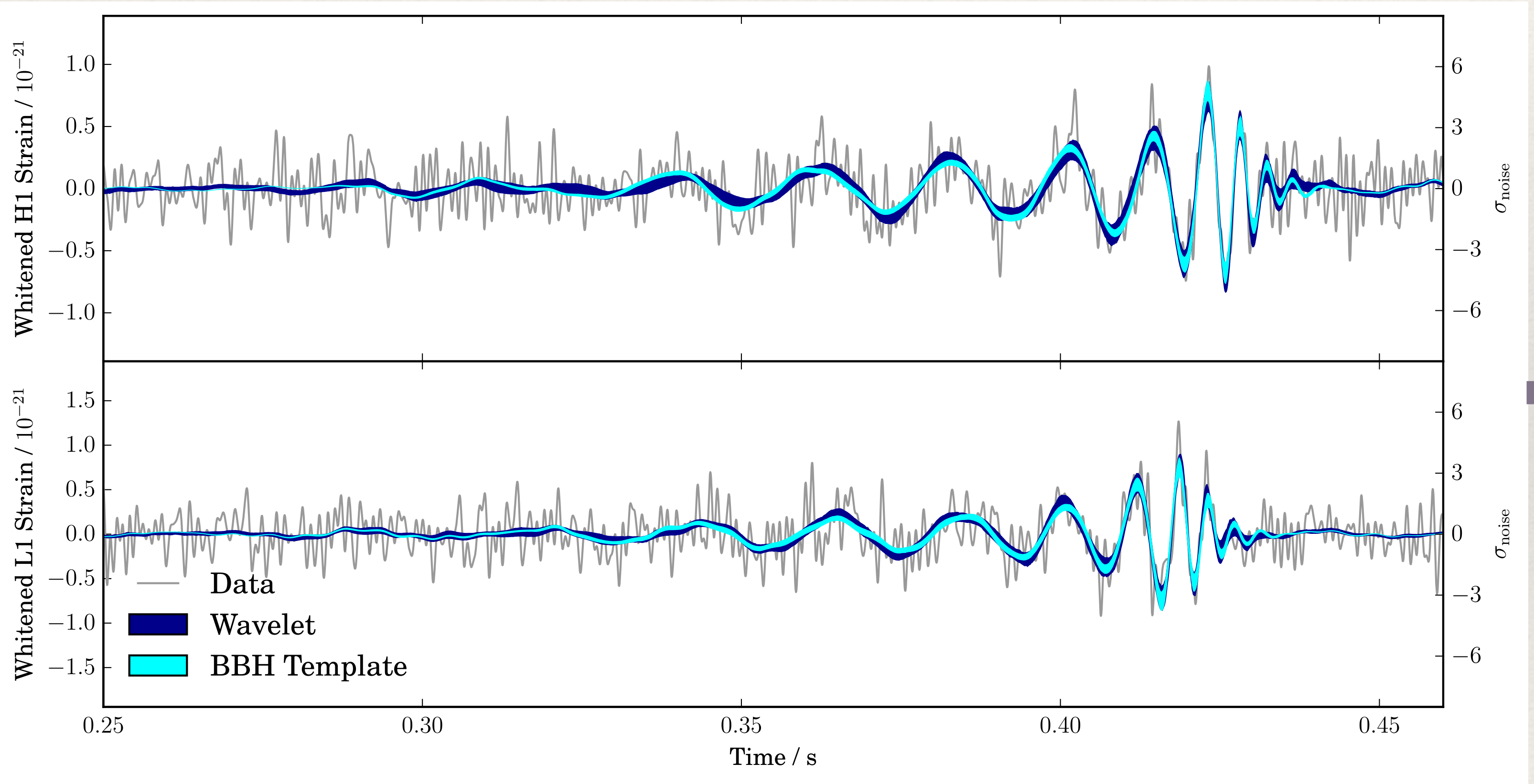
GW data analysis

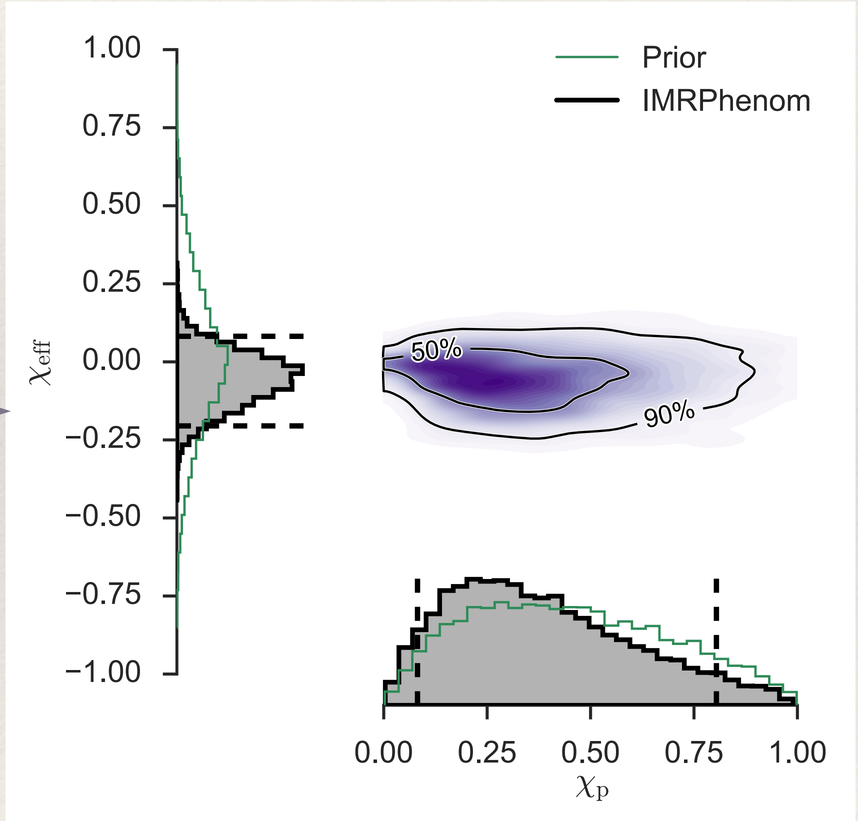


GW data analysis

$$p(\theta|d) = \frac{p(d|\theta)\ \ p(\theta)}{p(d)}$$

Fit to the data not a single line but multiple (for each parameter set in the posterioir)





[GW150914, LSC+VIRGO PRL (2016)]

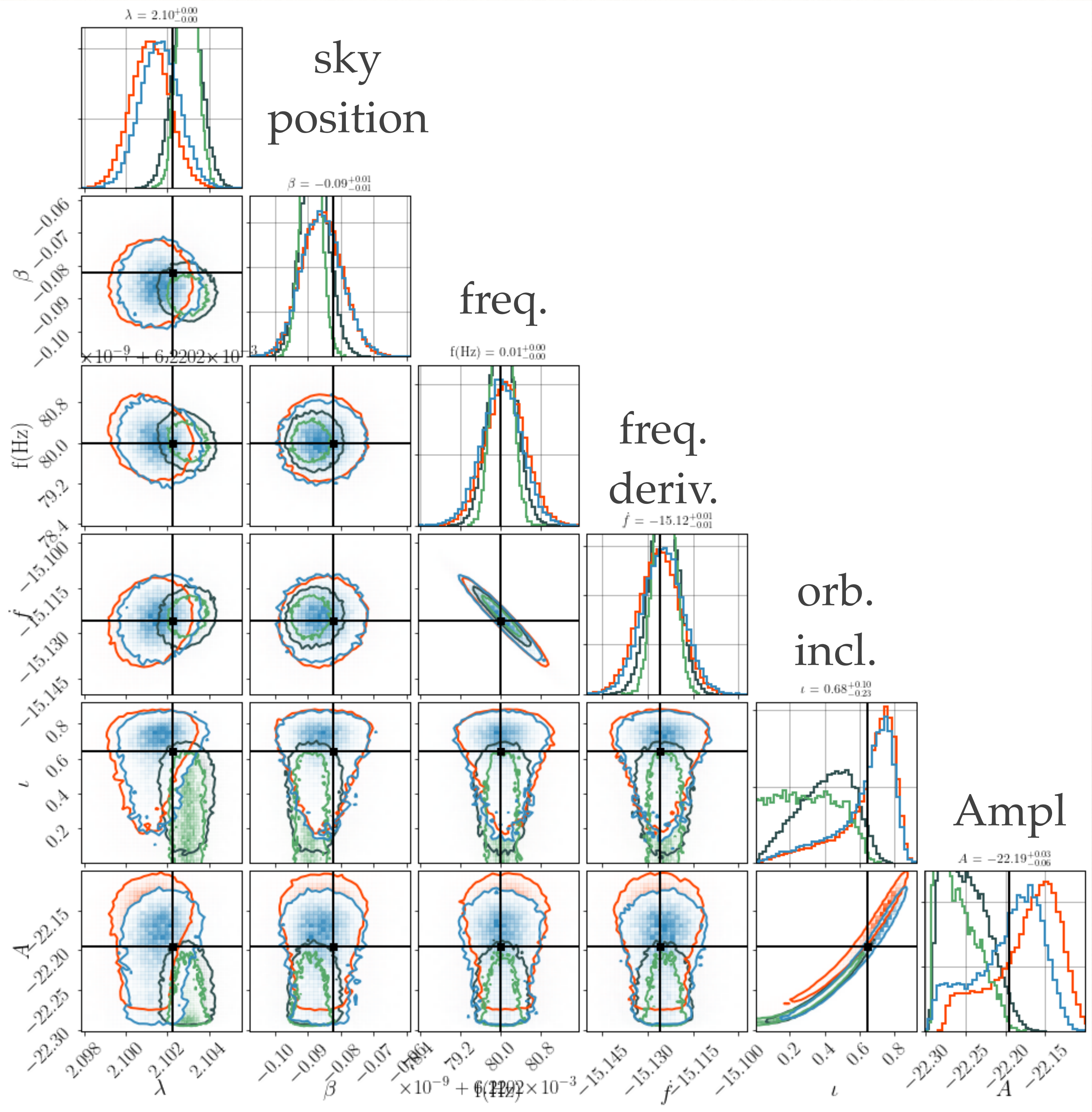
“LISA data challenges”

On going project within LISA consortium:

- Produce and publish simulated LISA data aiming at solving particular data analyis problem(s)
- Infrastructure which fixes conventions, LISA configuration, prototype for end-to-end LISA simulator and for data analysis pipeline
- The first data challenge “Radler” is over https://lisa-ldc.lal.in2p3.fr/ldc but you can still download data, analyze and submit results (recommended before moveing to the second challenge)
- Participants of “Radler challenge”: Birmingham University, Marshall (NASA), Montana Uni, Barcelona (ICE, CSIC, IEEC), CEA/IRFU (France), Goddard (NASA), APC (France), University of Trieste, IISER (India),   University of Minnesota, Imperial College London.
- Methods:
  - Various types of MCMC (MH, parallel tempering, reversible jump, slice, DE) with custom-made proposals, nested sampling (dynesty)
  - Penalizing for extra parameters (overfitting)
  - Non-parametric method, F-statistic
  - “Speed-up”: heterodyning, GPU, “fast response”

lisa

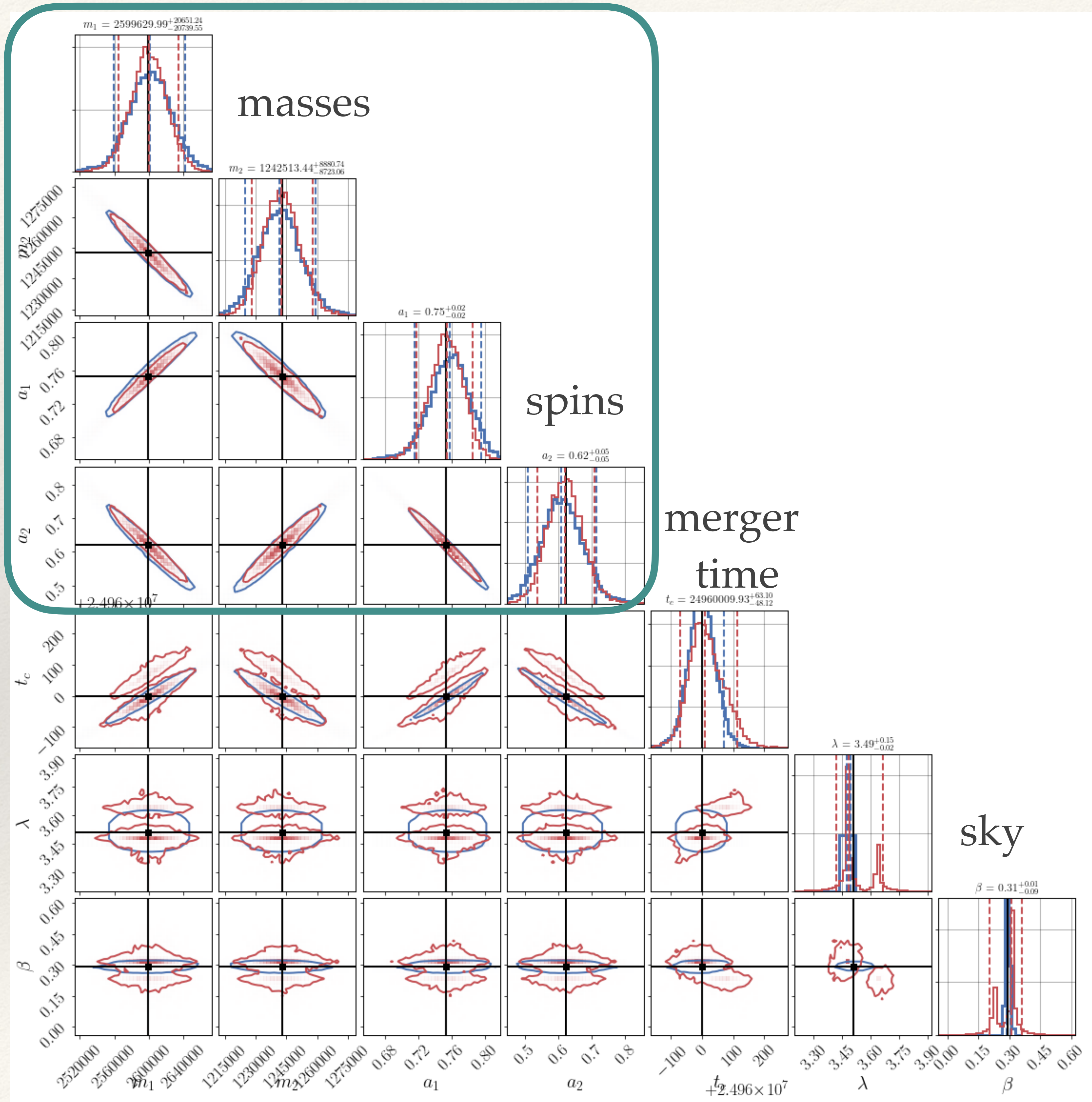
Analyzing verification Galactic binaries



- True parameters are within 90% credible region.
- Different priors are used
- Different noise realization v1, v2
- Different samplers

[Busciccio+ PRD, 2019]
[Littenberg+ PRD, 2020]

Search for MBHBs

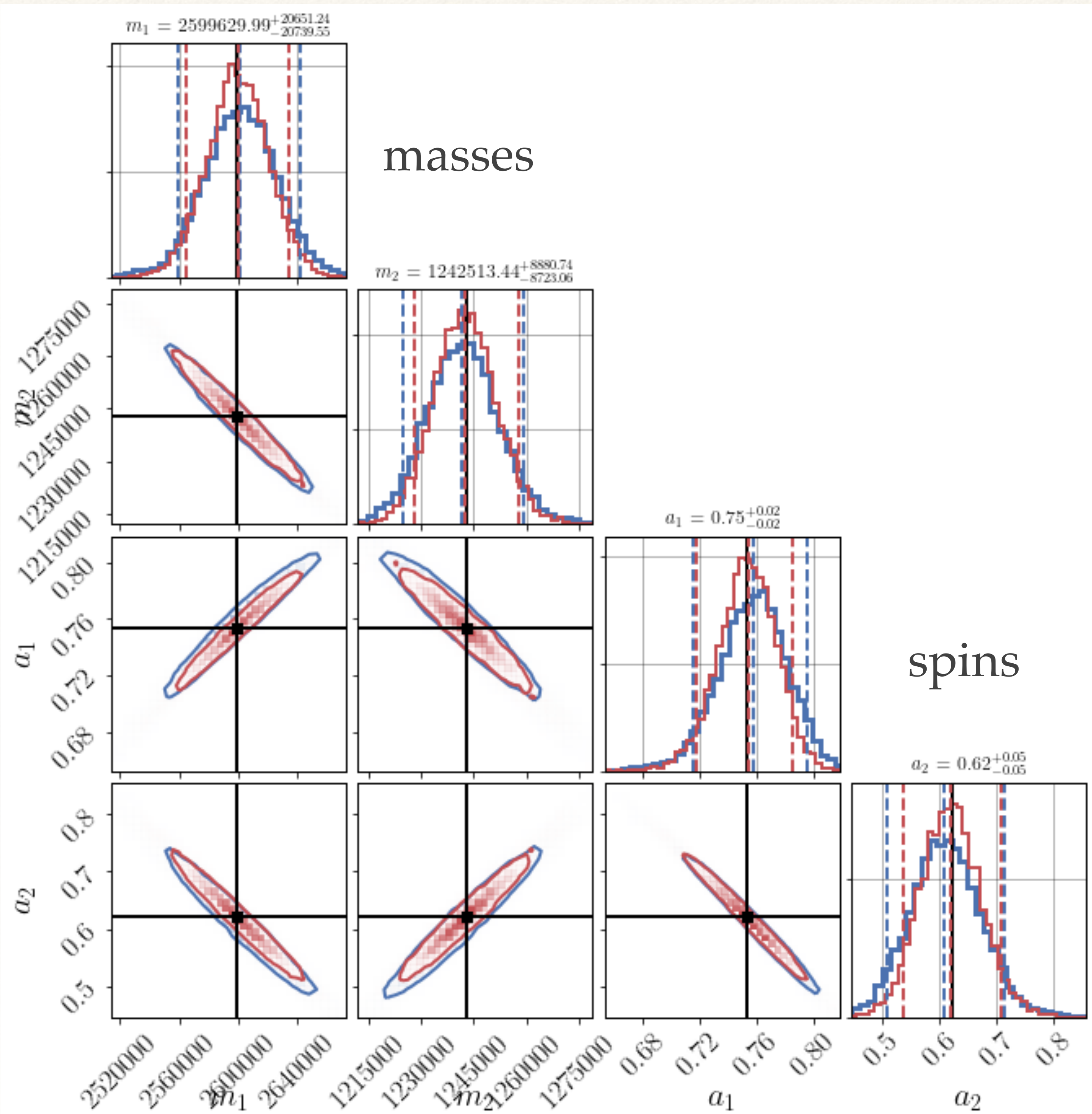


- GW signal includes inspiral merger and ringdown (no precession, only dominant mode)
- In total 11 parameters per source
- SNR ~400

[Cornish & Shuman PRD (2020)]
[Katz+ PRD (2020)]

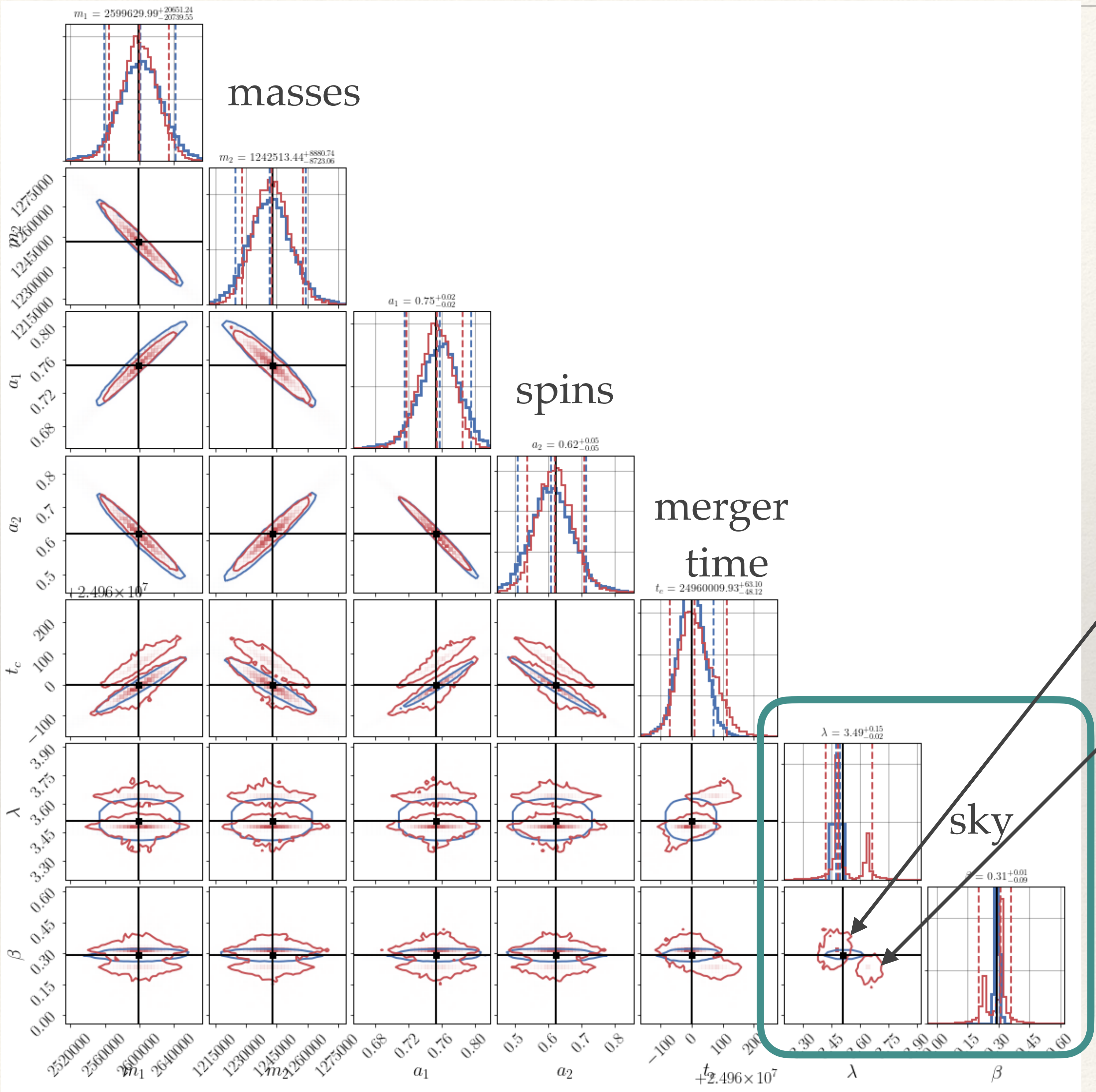
lisa

Search for MBHBs



Recover well "intrinsic parameters":
masses and spins of BHs

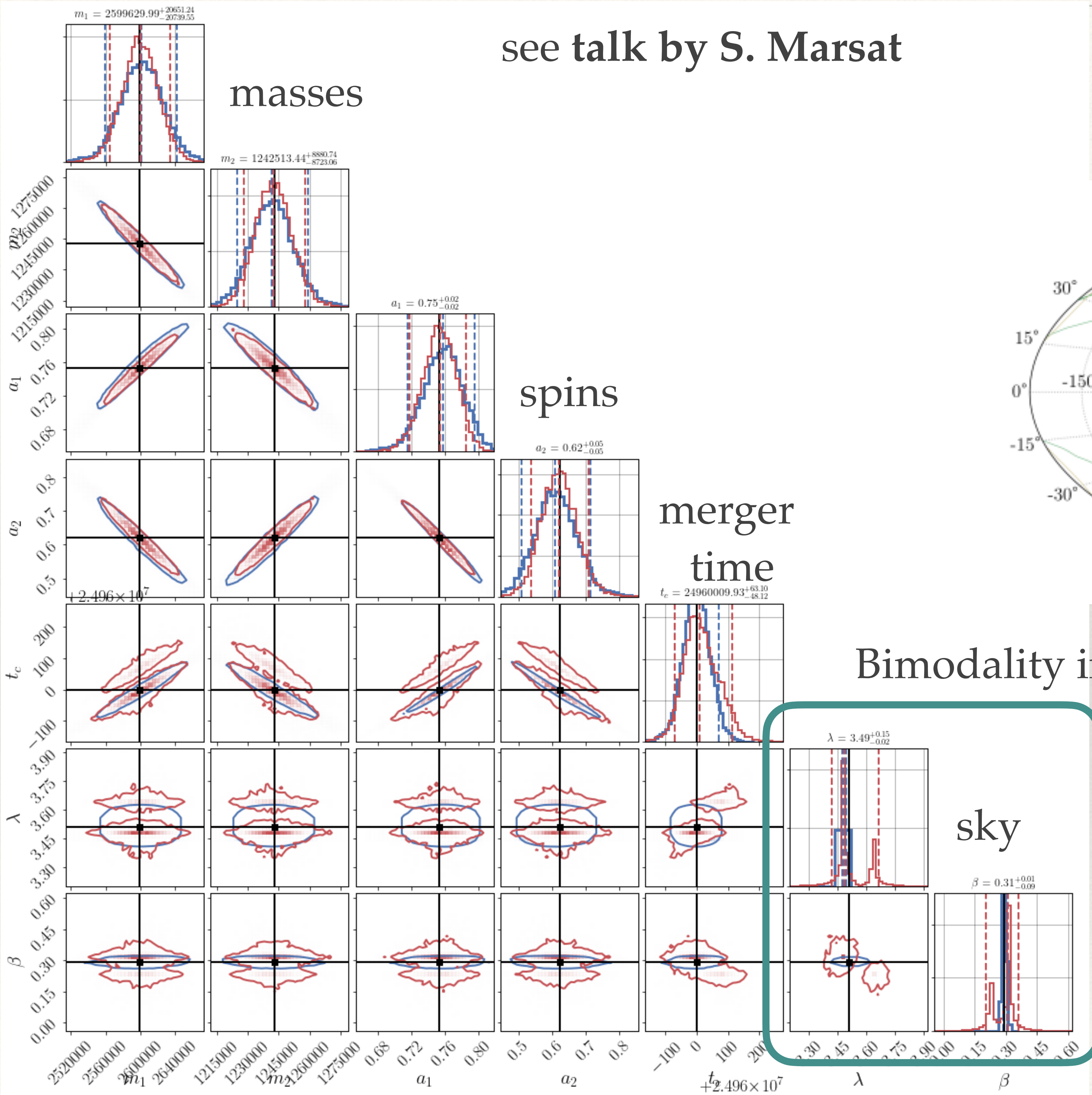
Search for MBHBs



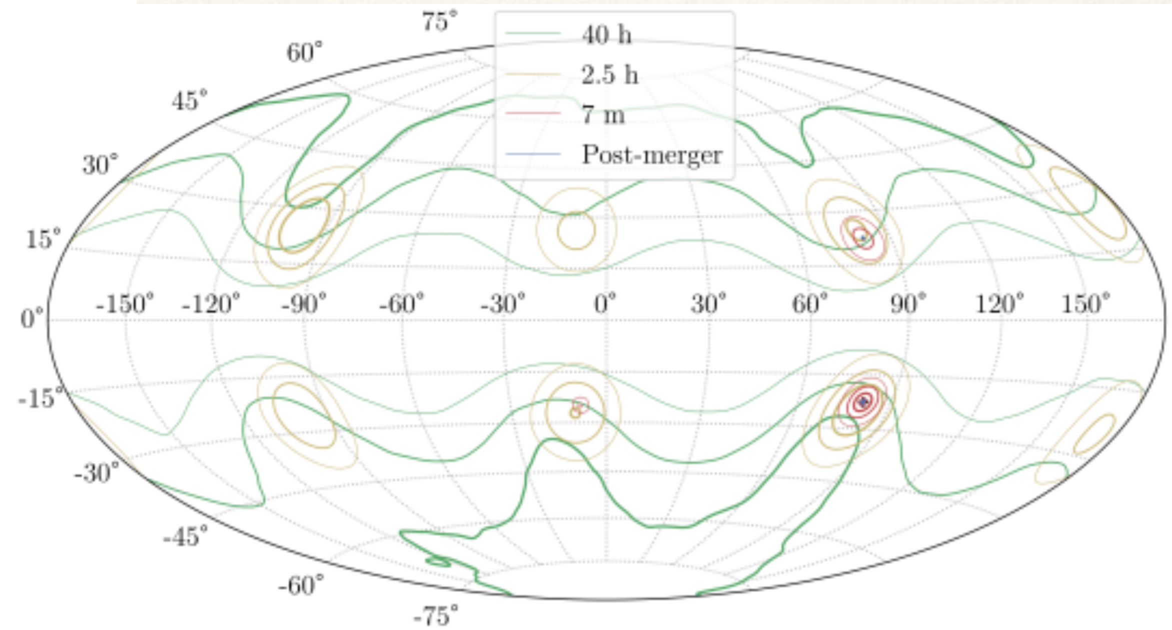
Bimodality in the position of source in the sky

Search for MBHBs

see **talk by S. Marsat**



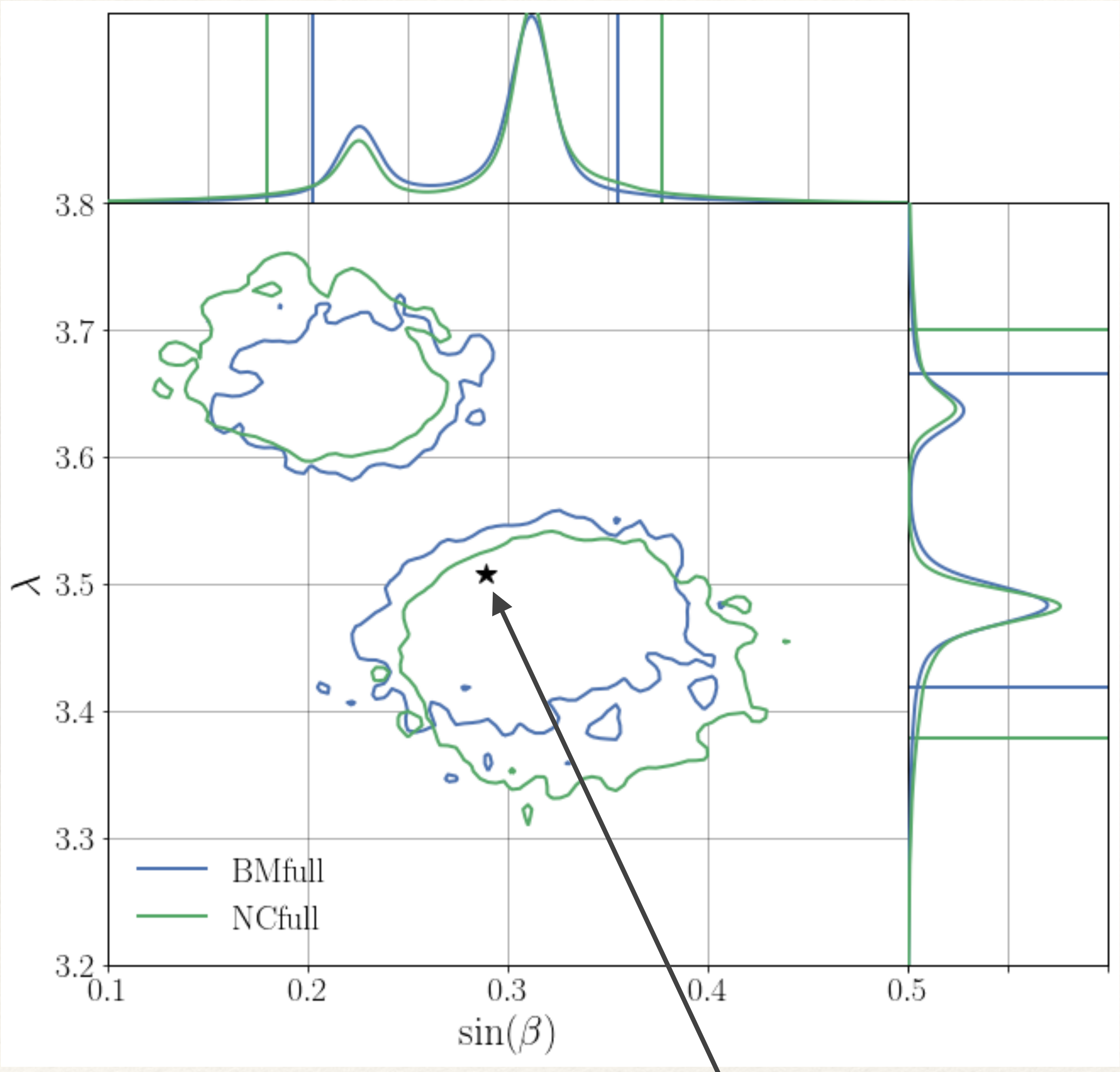
In general we could expect up to 8 local maxima in the sky



[Marsat+ arxiv 2003.00357]

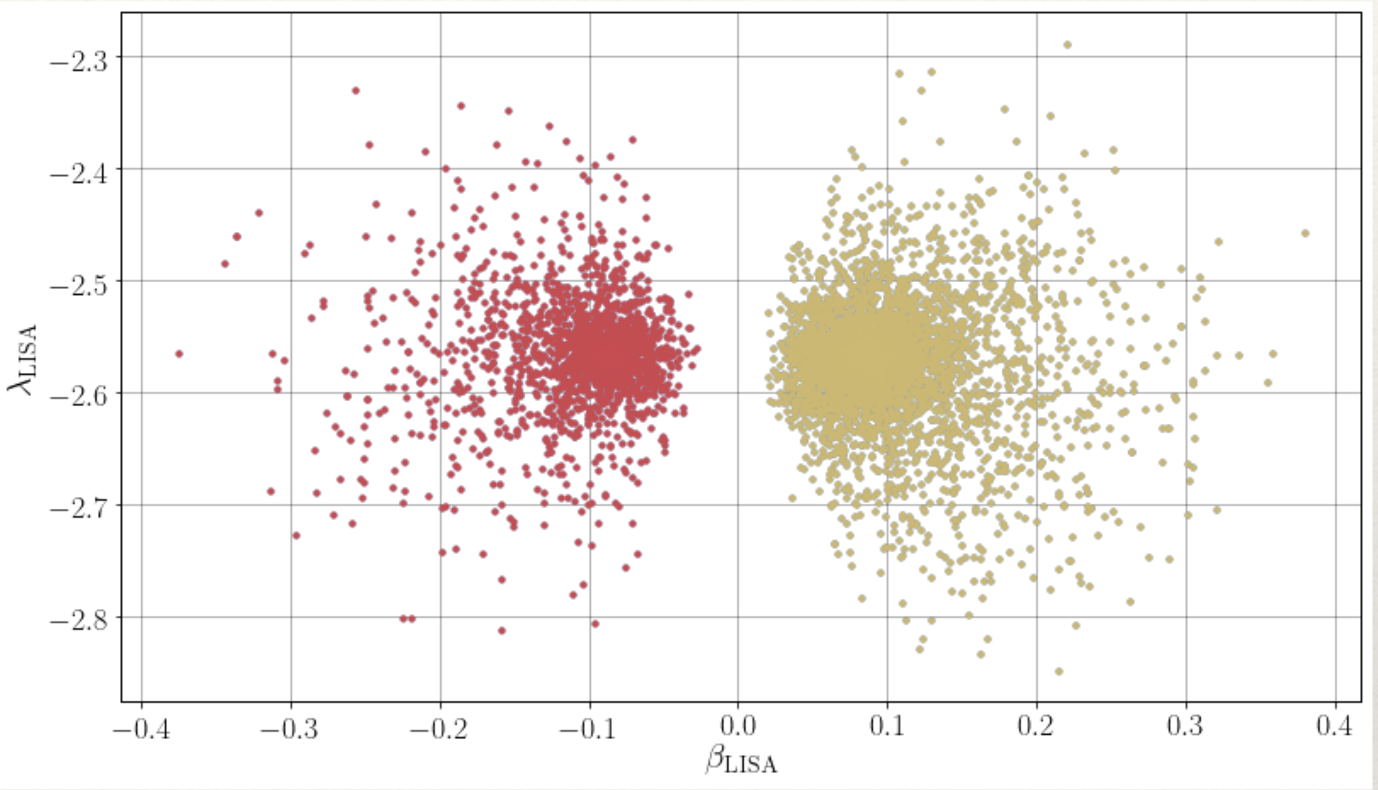
Understanding sky bi-modality

Sky 90% credible region in the solar system barycenter frame.
Amplitude ratio of two peaks ~ odd ratio



true position

Sky uncertainty in the LISA frame: symmetric w.r.t LISA's plane
The detectable part of signal: ~10 hours

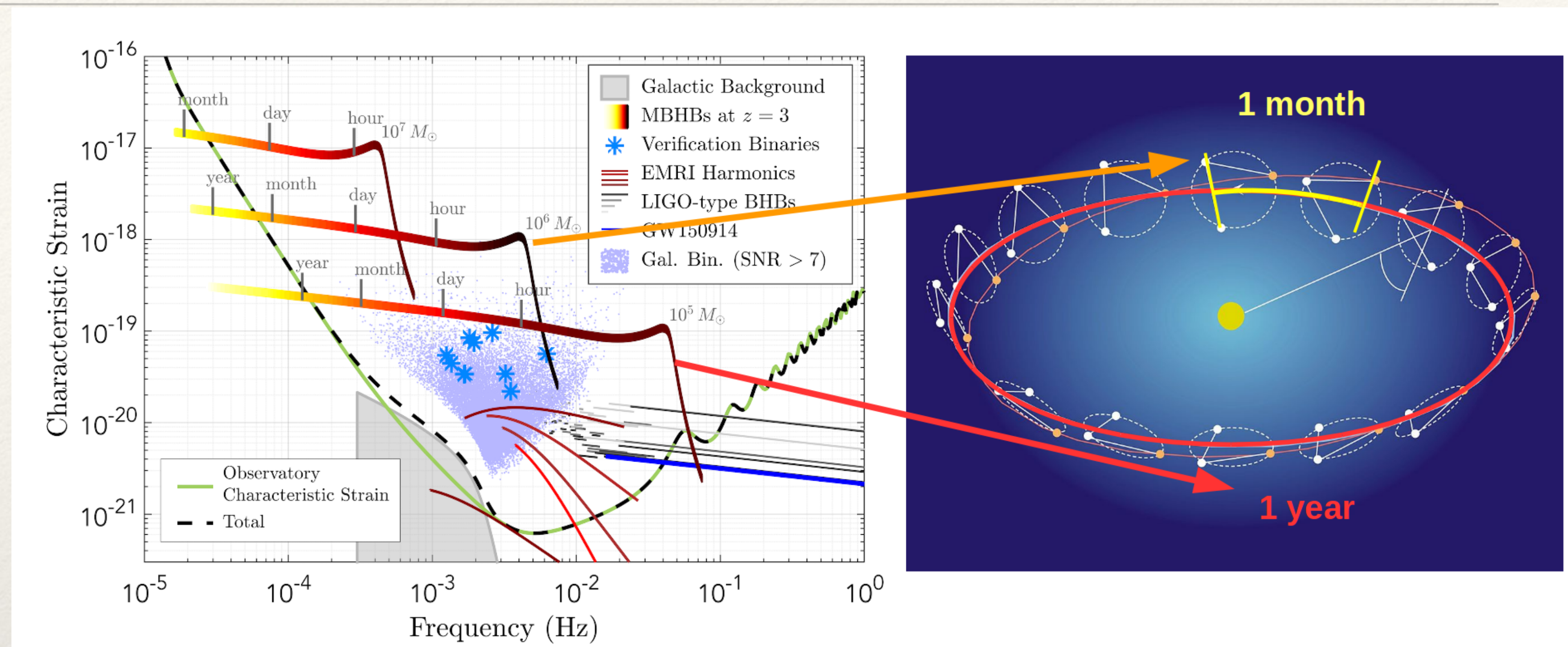


lisa

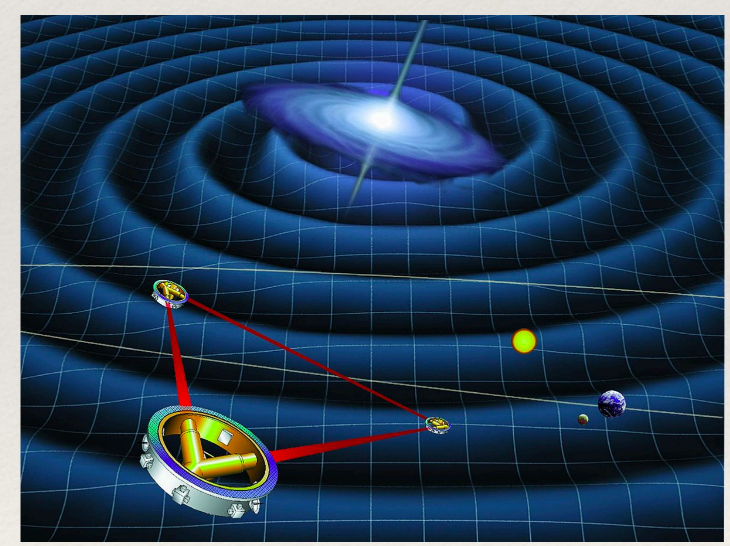
Understanding sky bi-modality

Doppler & amplitude modulation due to LISA's orbital motion

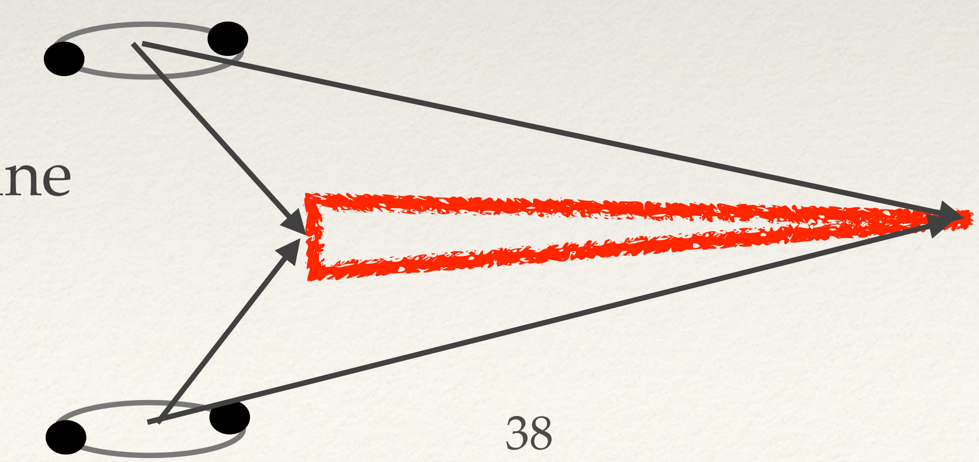
Direct and antipodal sky position (SSB frame)



Short duration: LISA's response — sensing difference in time arrival of GW at each spacecraft (small effect)



Symmetry wrt LISA's plane

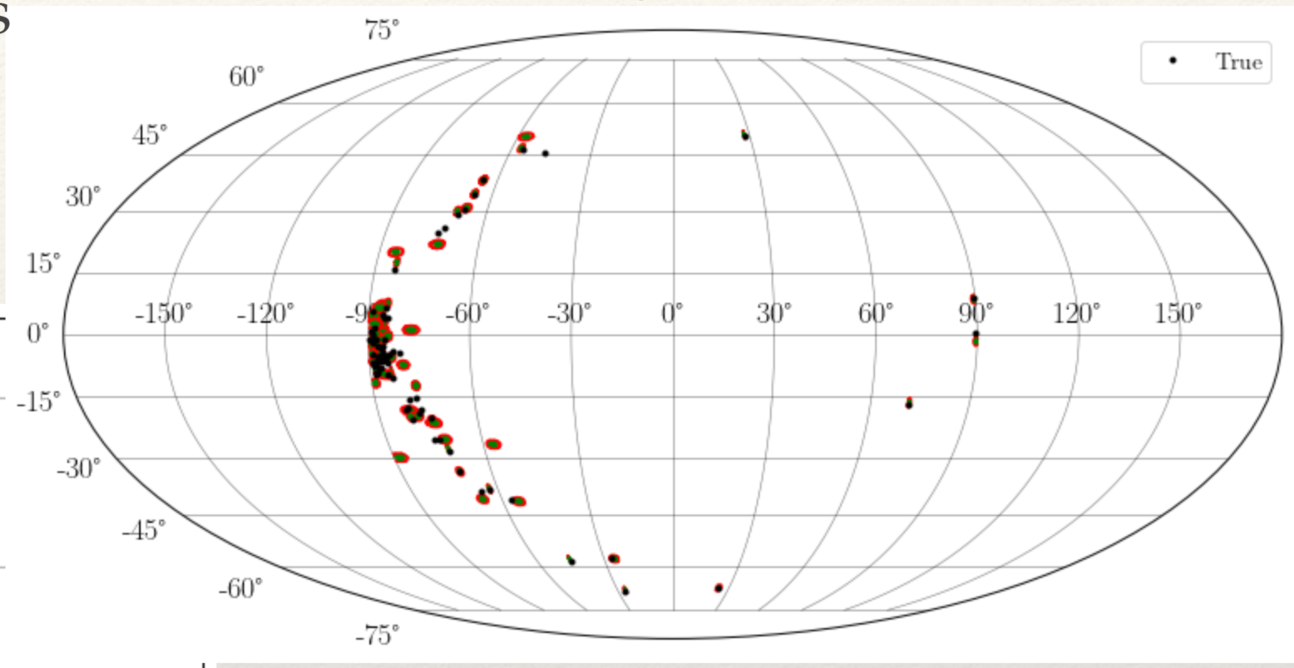


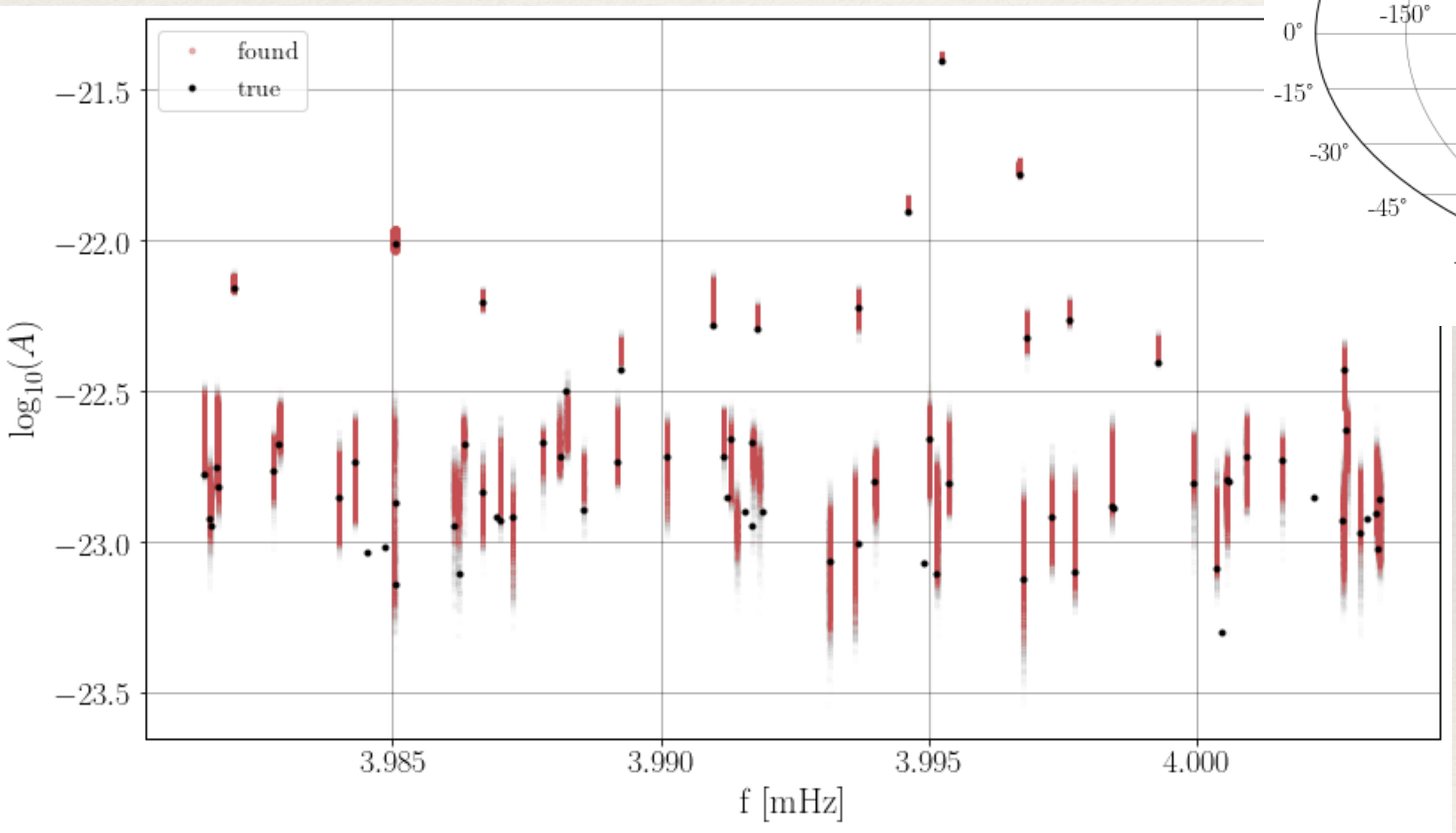
lisa

Galaxy (white dwarf binaries)

- The data contained 30 mln signals
- 10-20 thousands individually resolvable
- Results are submitted in small frequency bands
- Problem: to understand the number of sources

[Littenberg+ PRD, 2020]

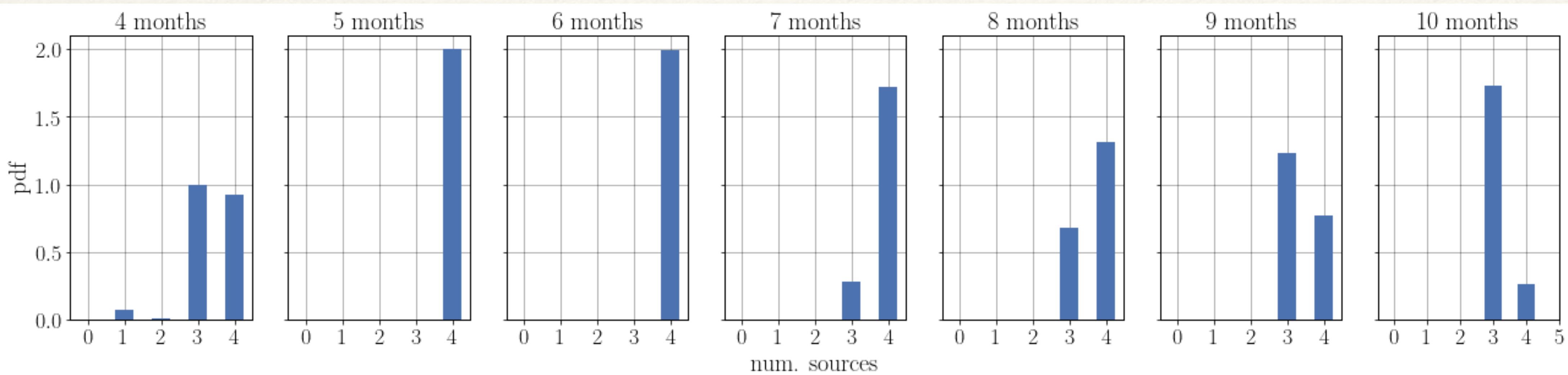




lisa

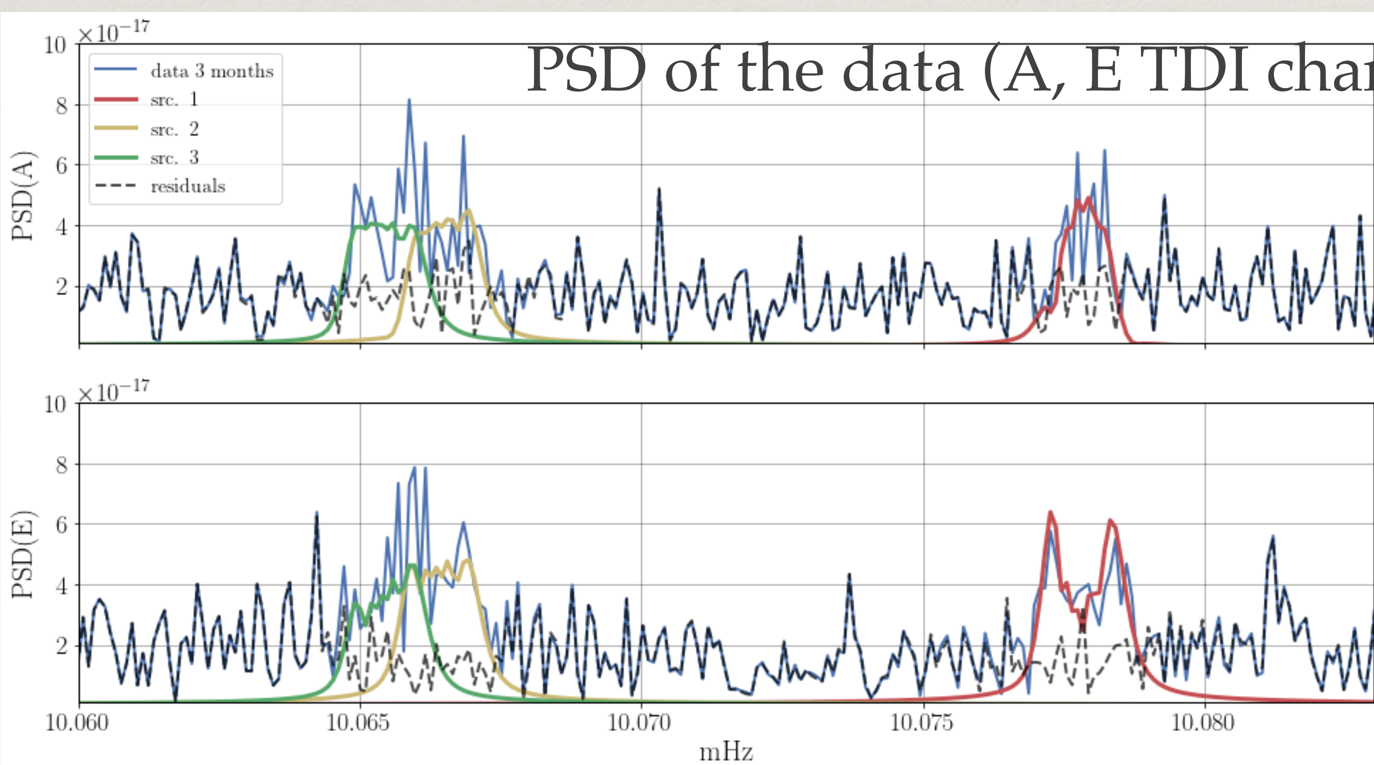
Identifying number of sources

- Take a narrow band next to 10mHz: 3 GW signals
- Perform time-data adaptive (4 months, 5 months, … 10 months) search
- Consider 4 models: 1 GW source, 2 GW sources, 3 GW source, 4 GW sources



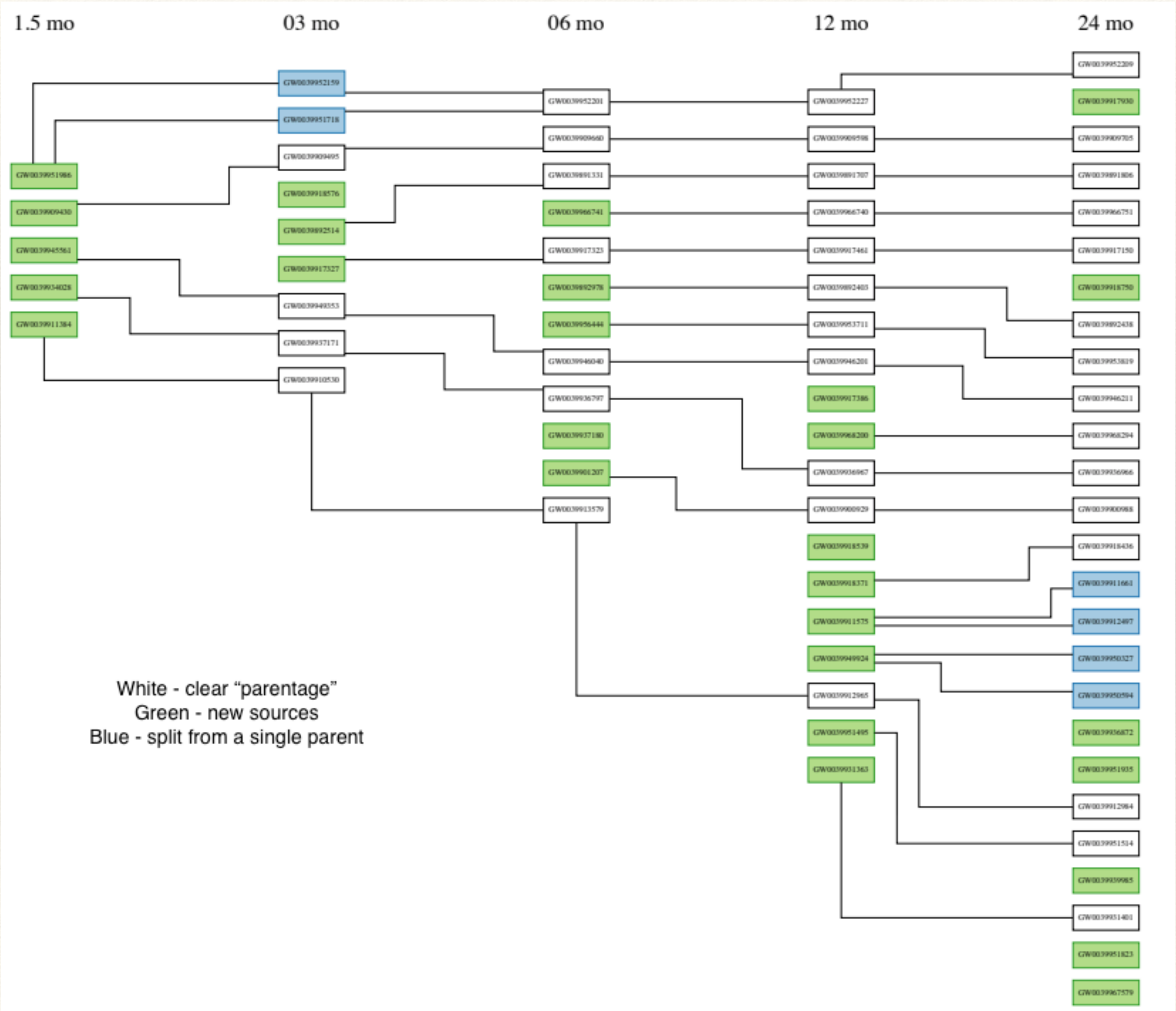
Probability of having *N*-sources in the data using different data duration

PSD of the data (A, E TDI channels)



lisa

Time-evolving catalogue building

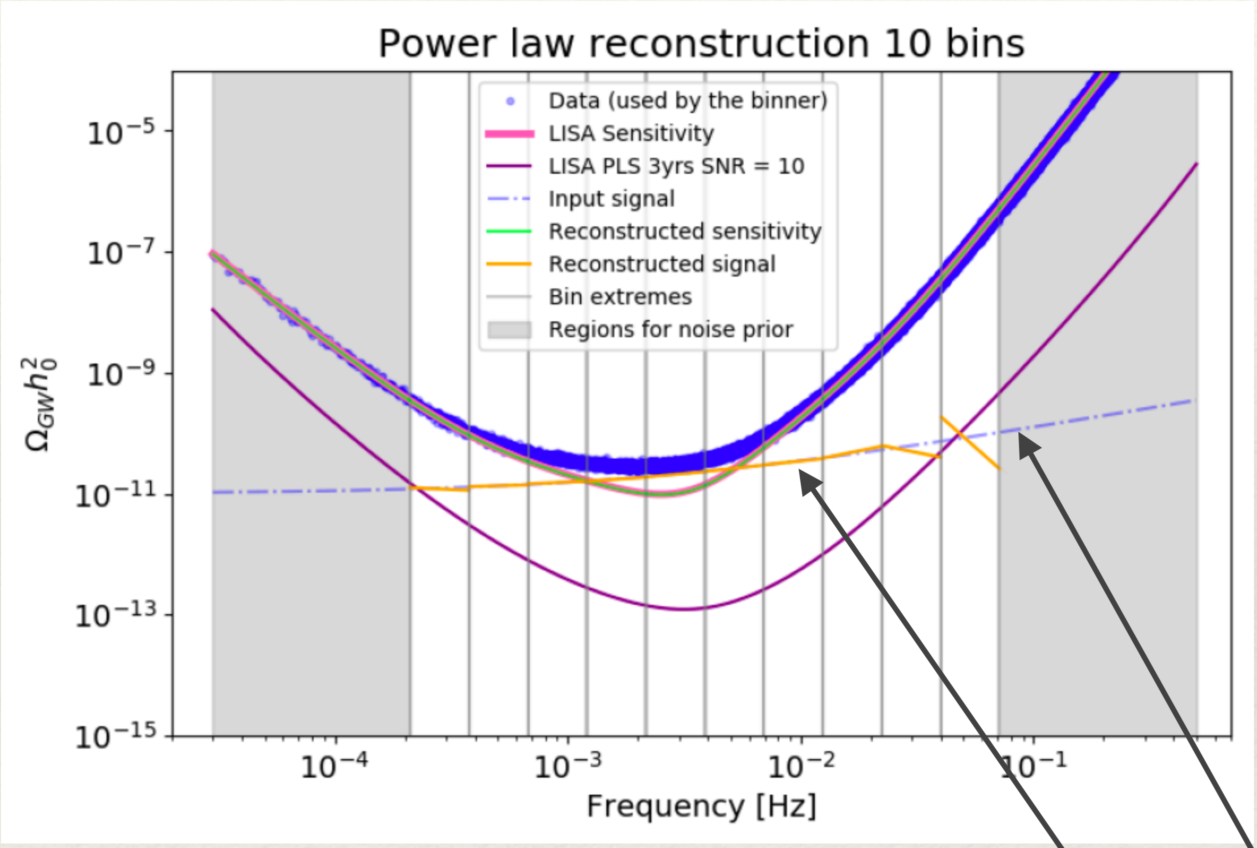


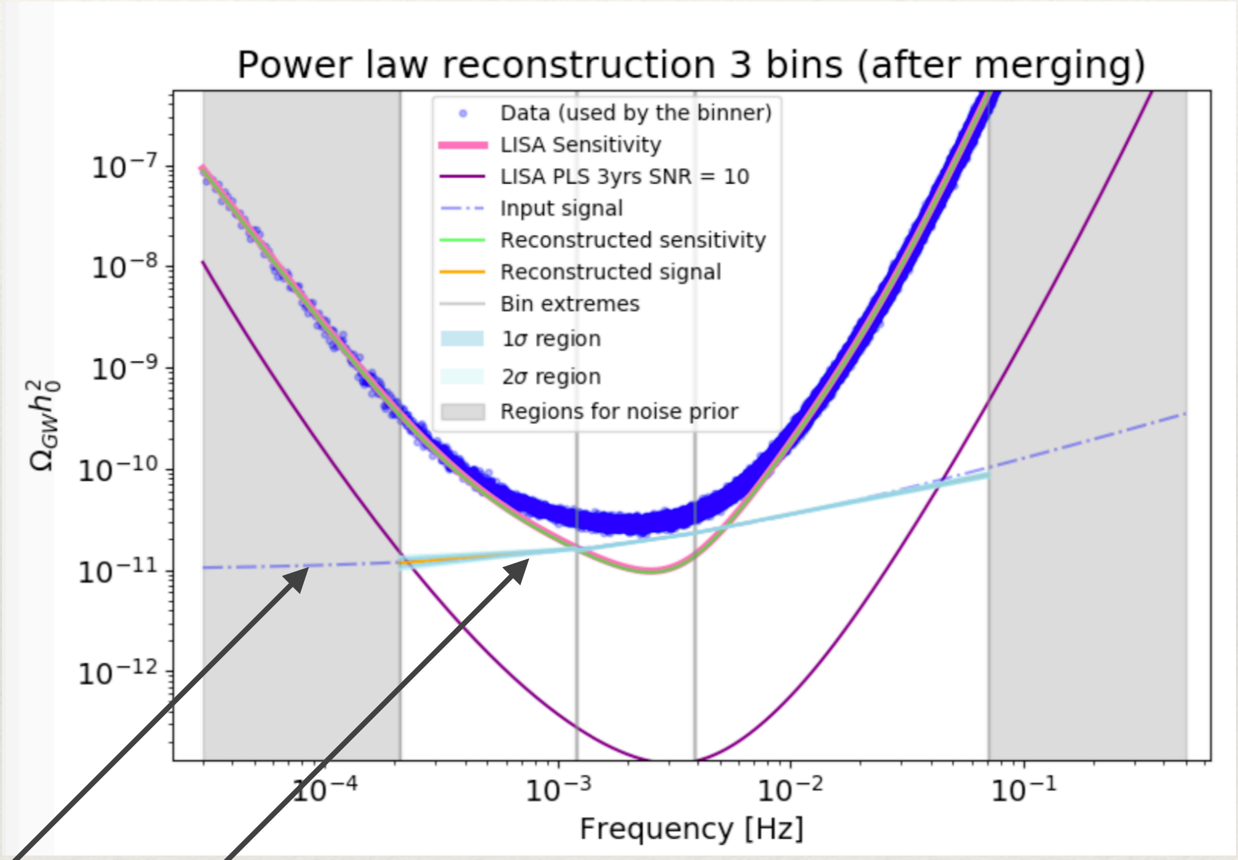
[Littenberg+ PRD, 2020]

Stochastic GW signal

Raphael Flauger, Nikolaos Karnesis, Germano Nardini, Mauro Pieroni, Angelo Ricciardone, Jesus Torrado, Ch. Caprini, Sharan Banagiri, Alexander Criswell, Vuk Mandic

split freq. in bins and evaluate SGW piecewise (powerlaw). Penalize for unnecessary large number of bins (AIC)



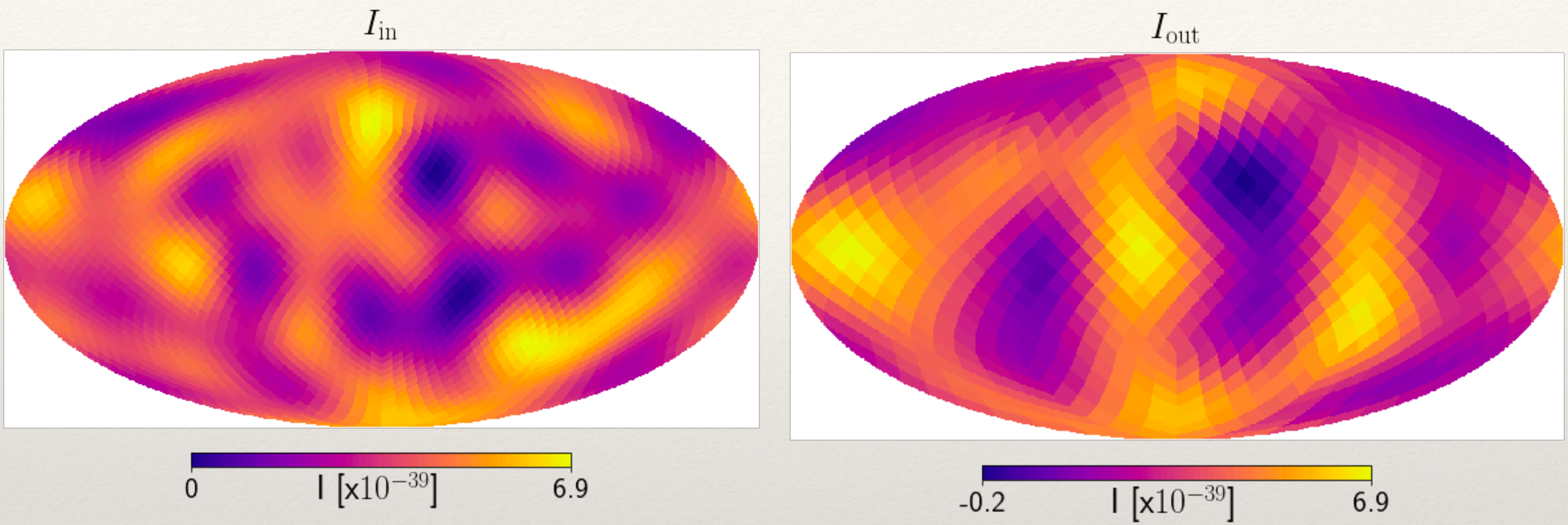


Injected

Reconstructed

Stochastic GW signal

Reconstruction of Anysotropy

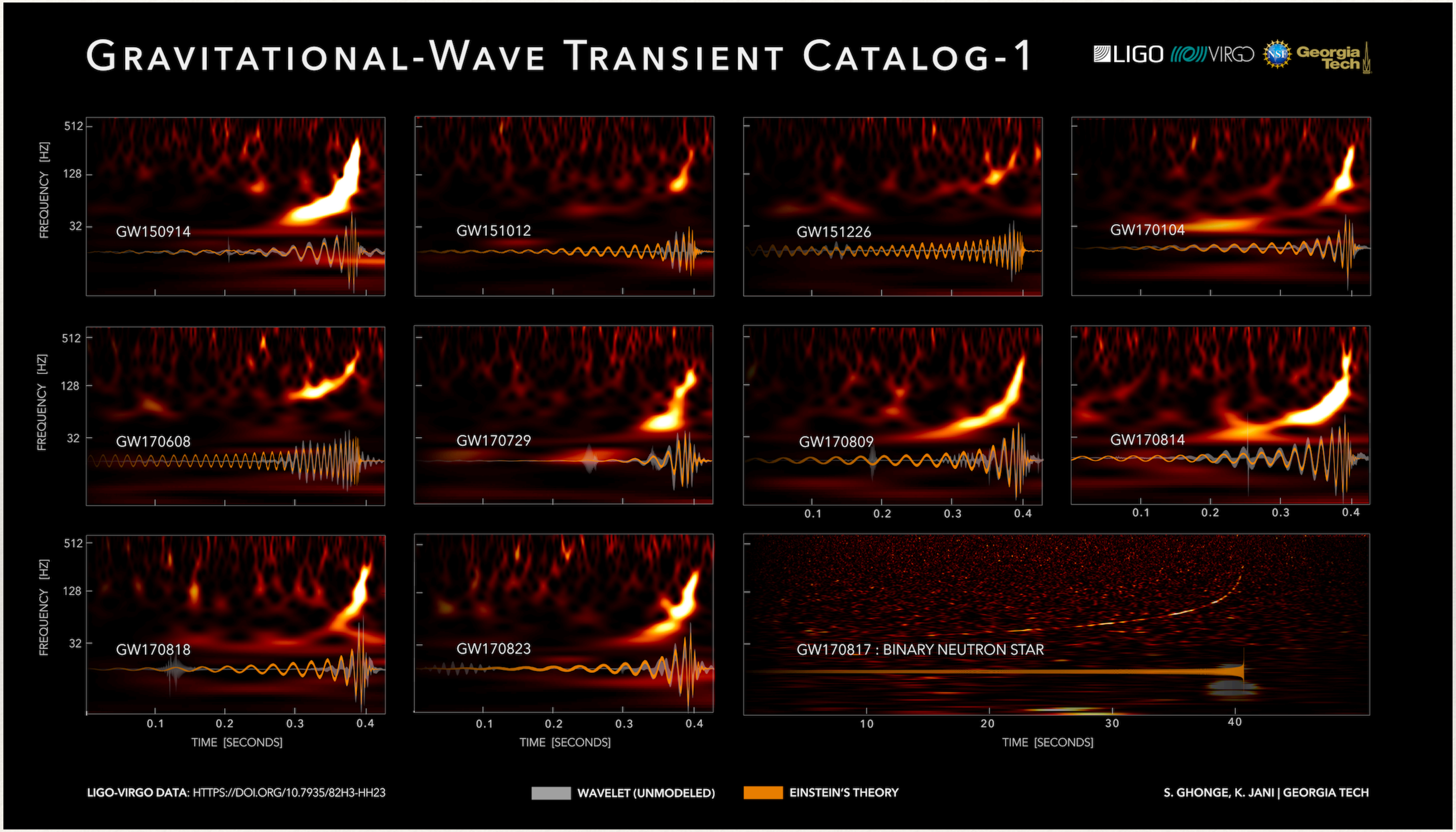


Smoothing on small scales (complete loss of sensitivity at l~ 15)

[Contaldi+ (2006)]

lisa

Stellar mass BH binaries

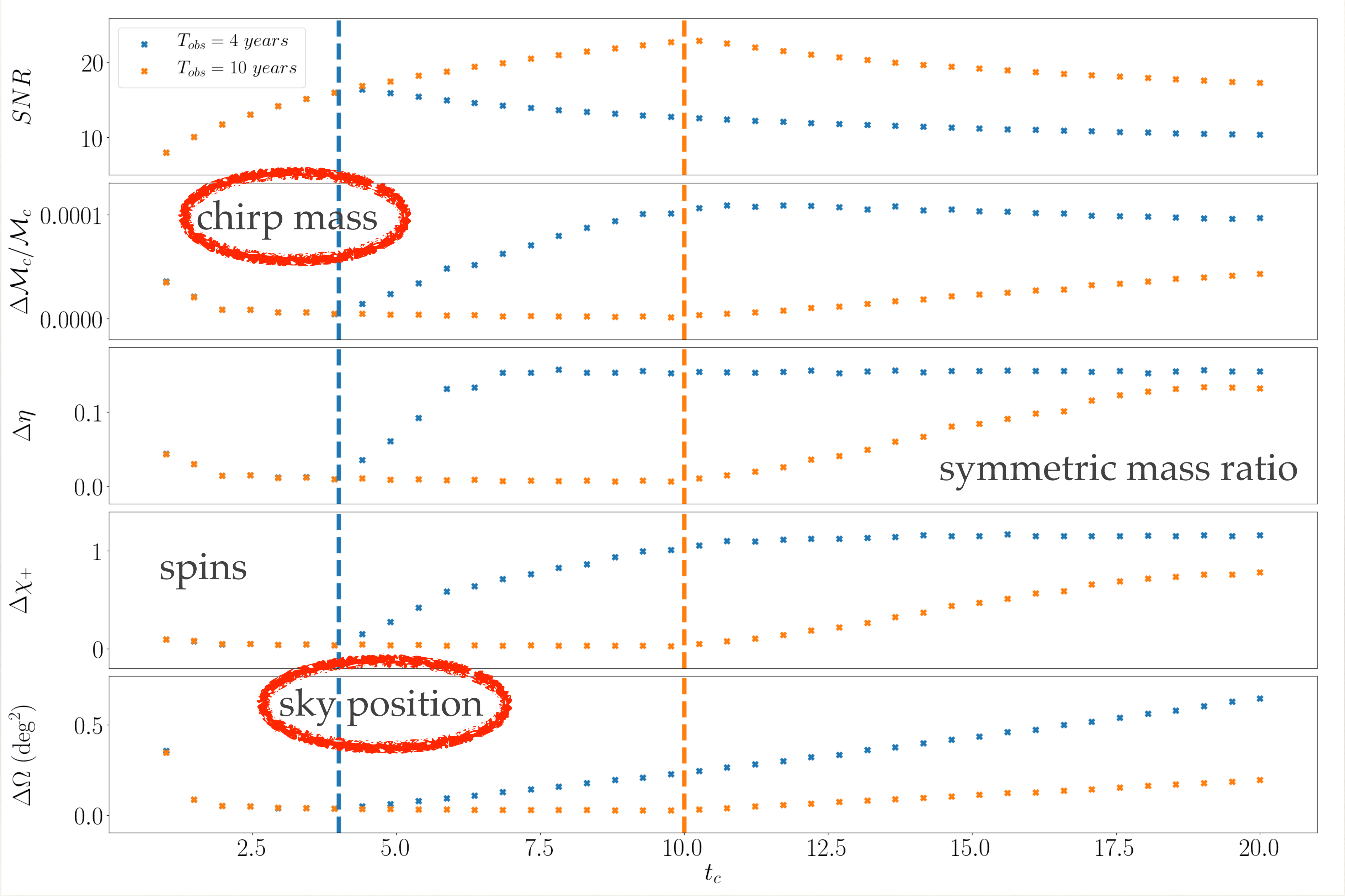


[LVC (2018)]

lisa

Stellar mass BH binaries

- Weak source at high frequency end of LISA' s sensitivity: possible multi-band sources
- Hard to detect. Could serve as laboratories for testing GR



Moved to challenge 1b

[Toubiana+ (2020), 2007.08544]
[Toubiana+, PRD (2020)]
[Sayantani Datta+, (2006) 2006.12137]
[Moore, Gerosa, Klein, MNRAS (2019)]

lisa

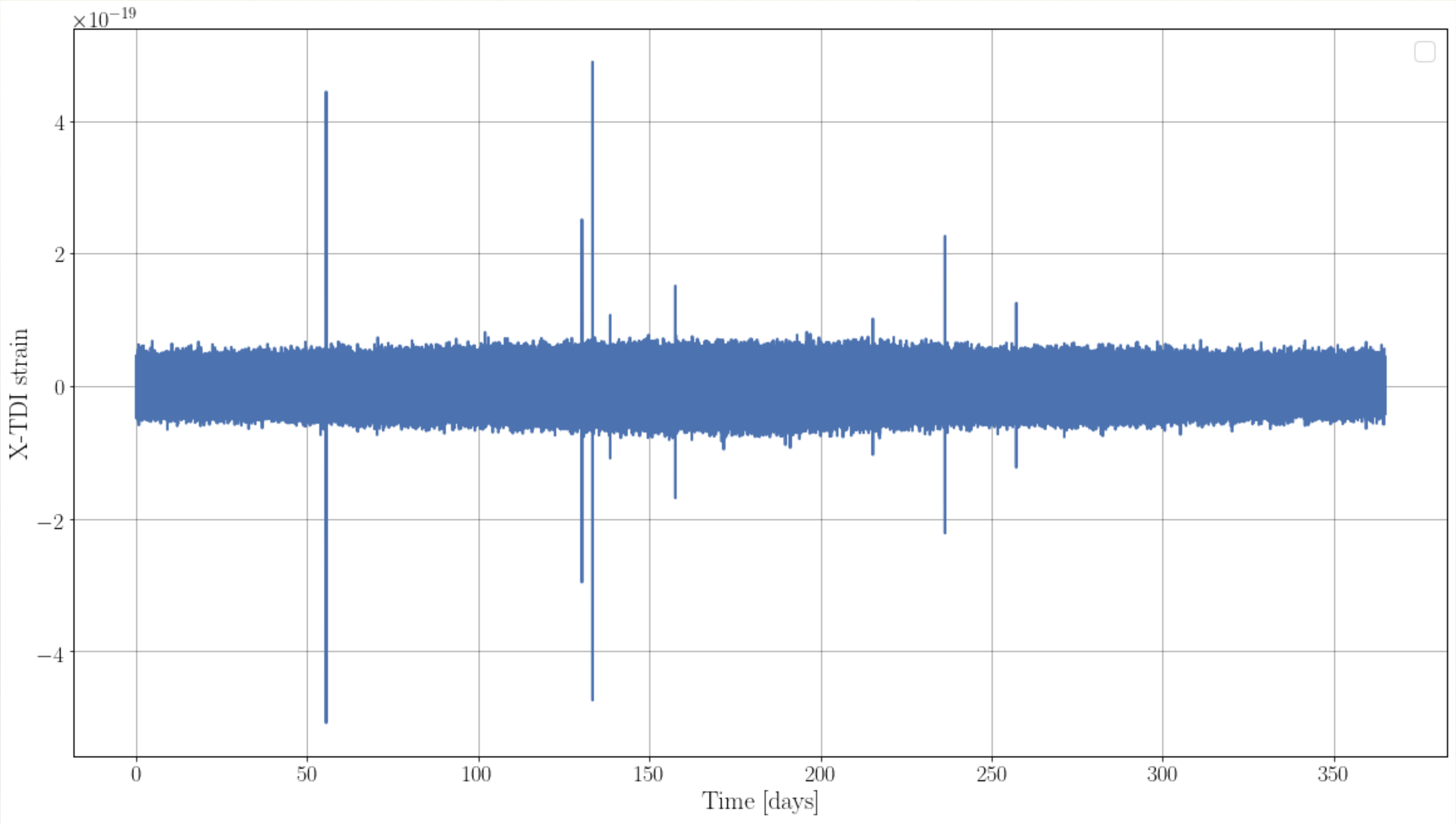
LISA data challenge-2 (“Sangria”)

Sangria (English: /sænˈɡriːə/, Portuguese pronunciation: [sɐ̃ˈɡɾi.ɐ]; Spanish: *sangría* [saŋˈɡɾi.a]) is an alcoholic beverage of Spanish origin. A punch, the sangria traditionally consists of red wine and chopped fruit, often with other ingredients such as orange juice or brandy.

- Mild “enchilada” data challenge:
  mixture of sources + unknow level of instrumental noise
- Galactic WD binaries (again 30 mln) plus unknown number of MBHBs:
  - Training data challenge + Blind challenge
  - Noise is not stationary (cyclo-stationary) due to stochastic foreground confusion noise
  - The data is available https://lisa-ldc.lal.in2p3.fr/ldc
  - Officially released

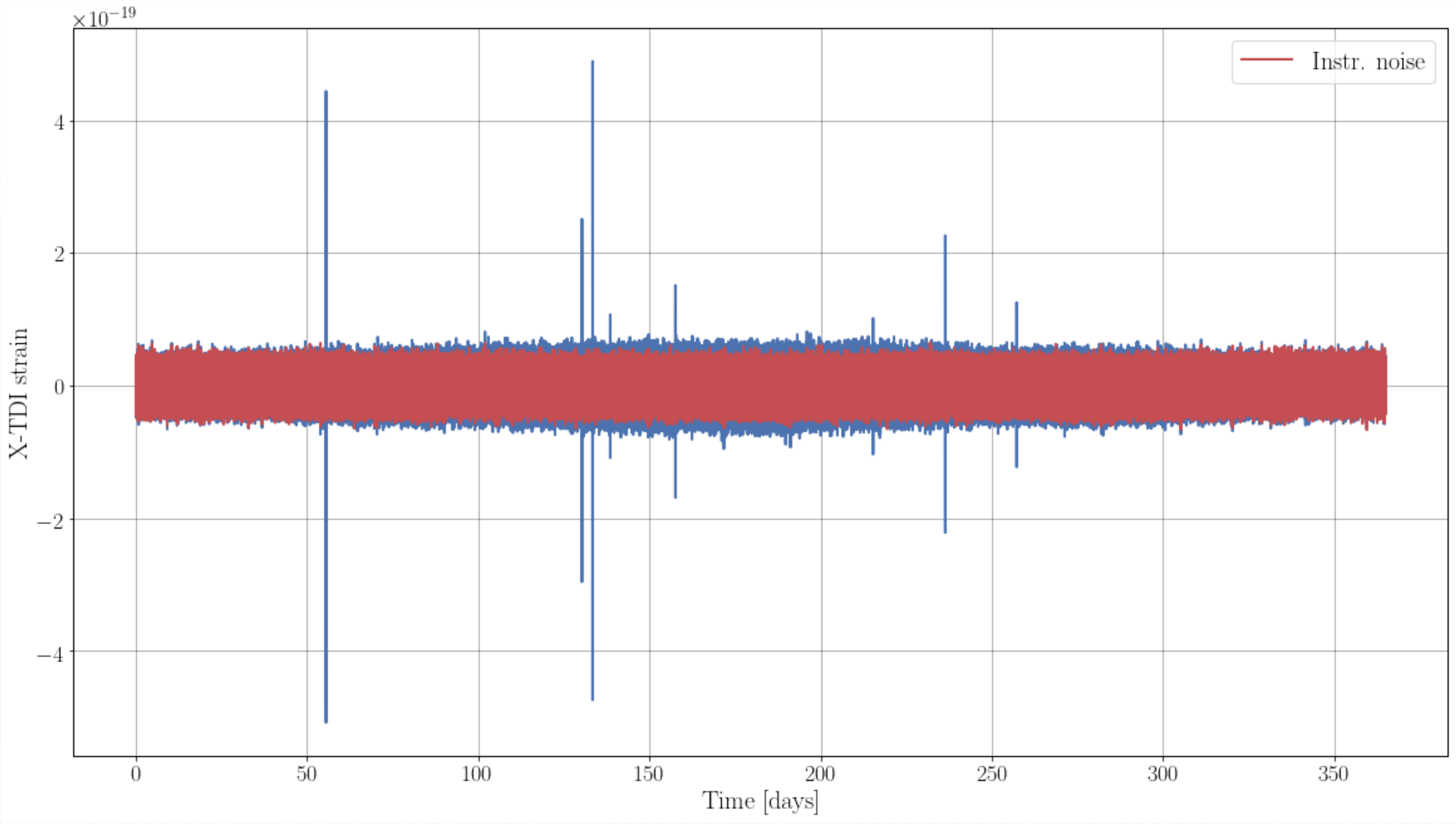
lisa

"Sangria" in time domain



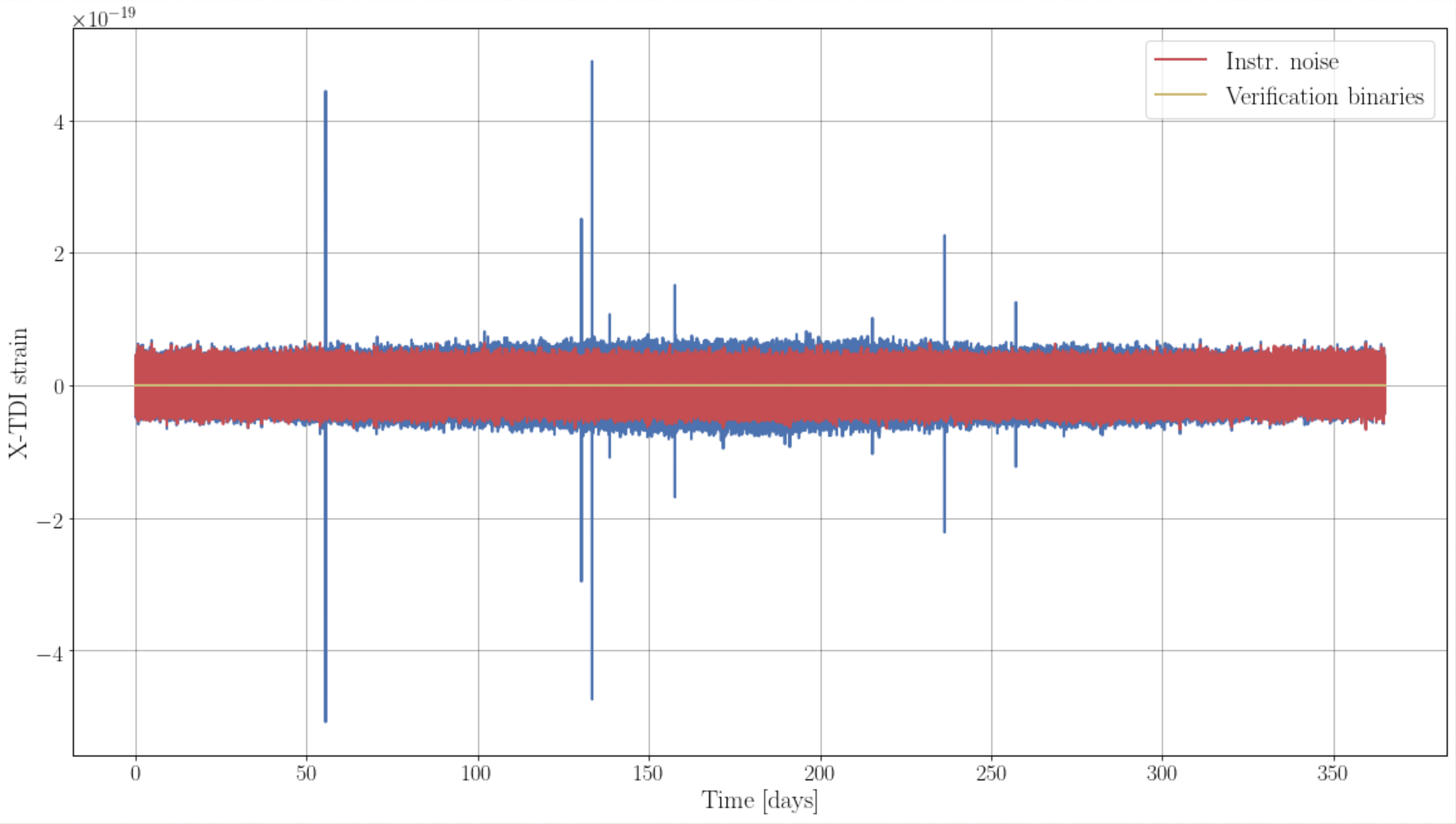
lisa

“Sangria” in time domain



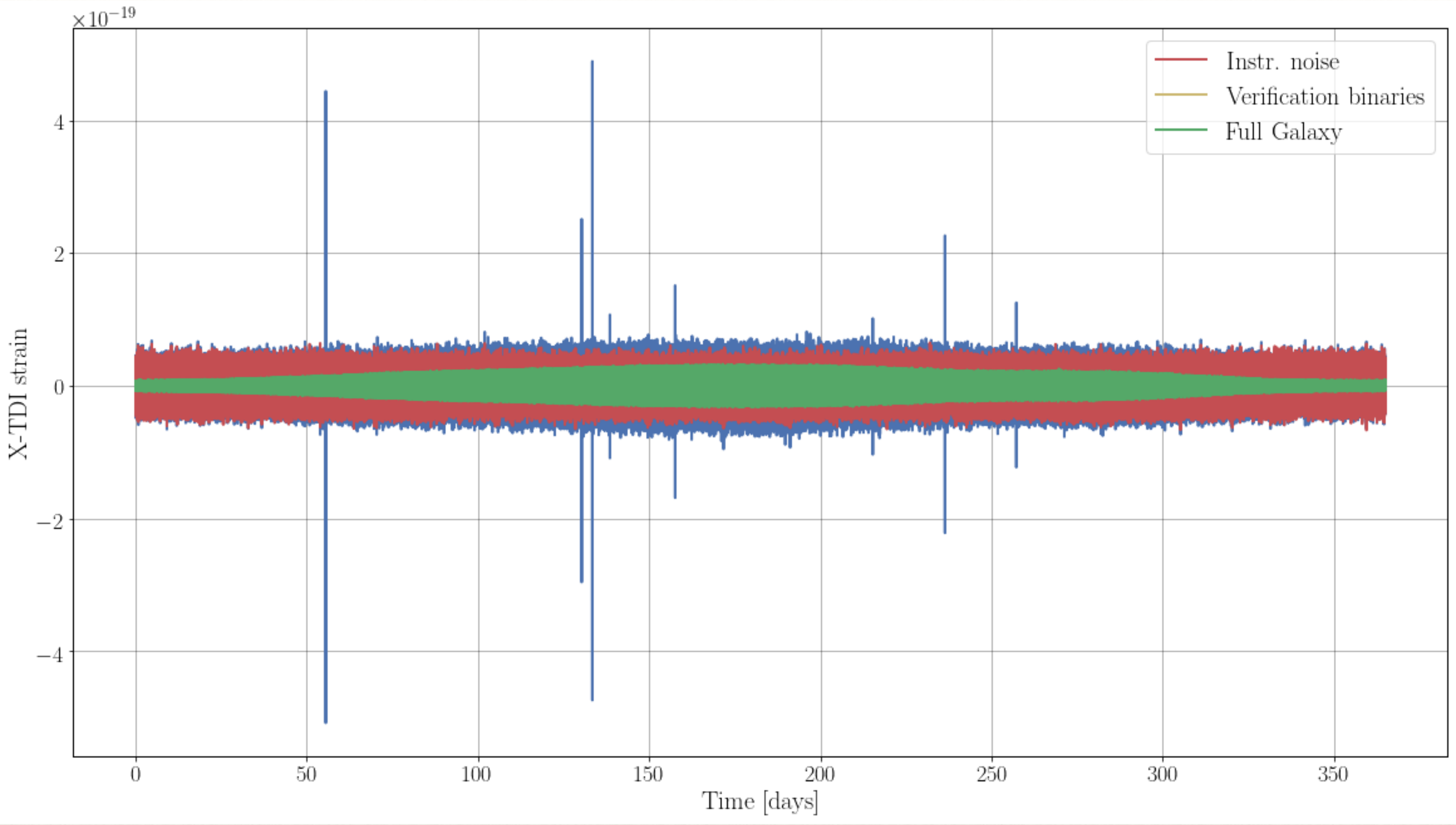
lisa

“Sangria” in time domain

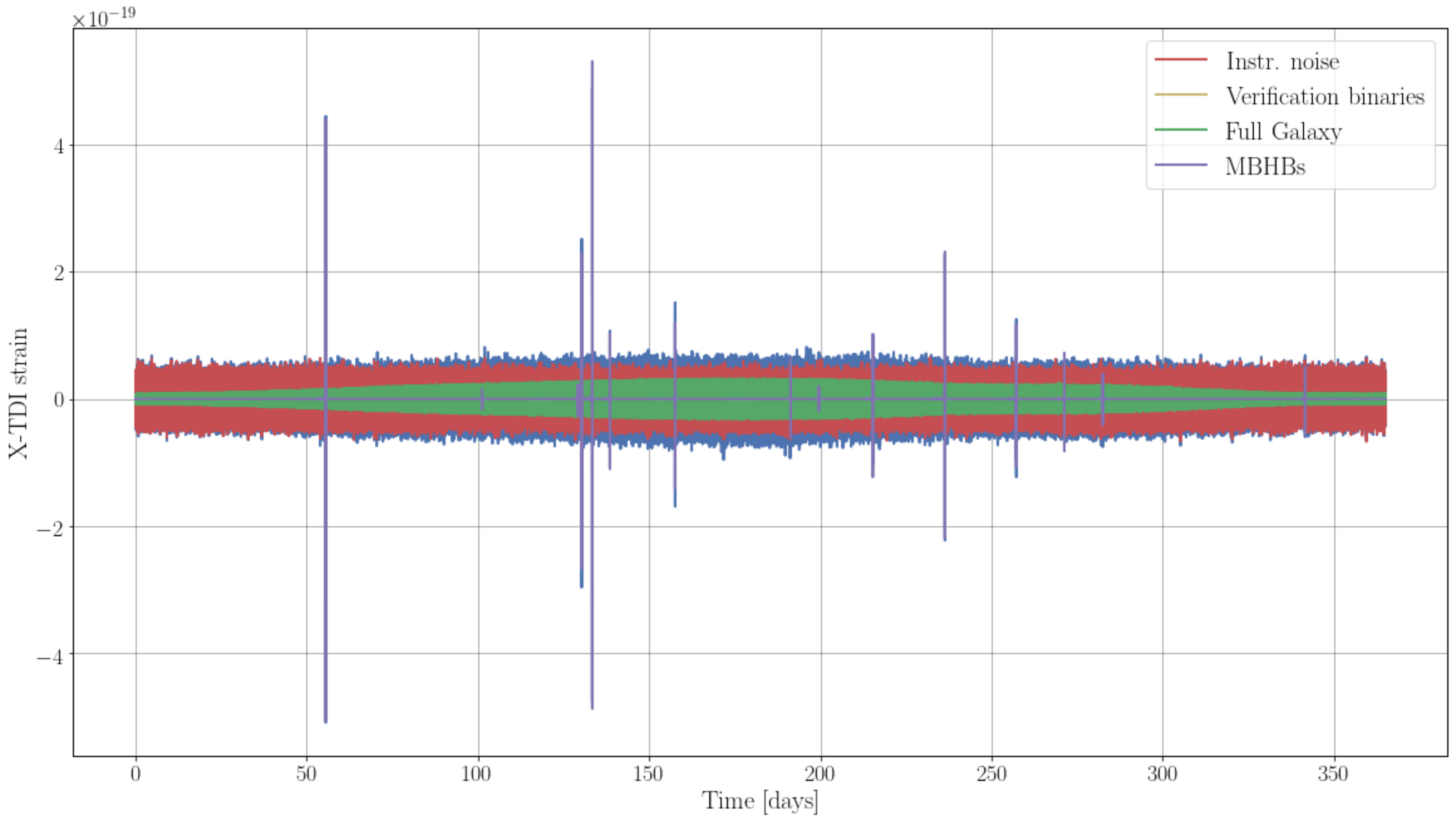


lisa

“Sangria” in time domain

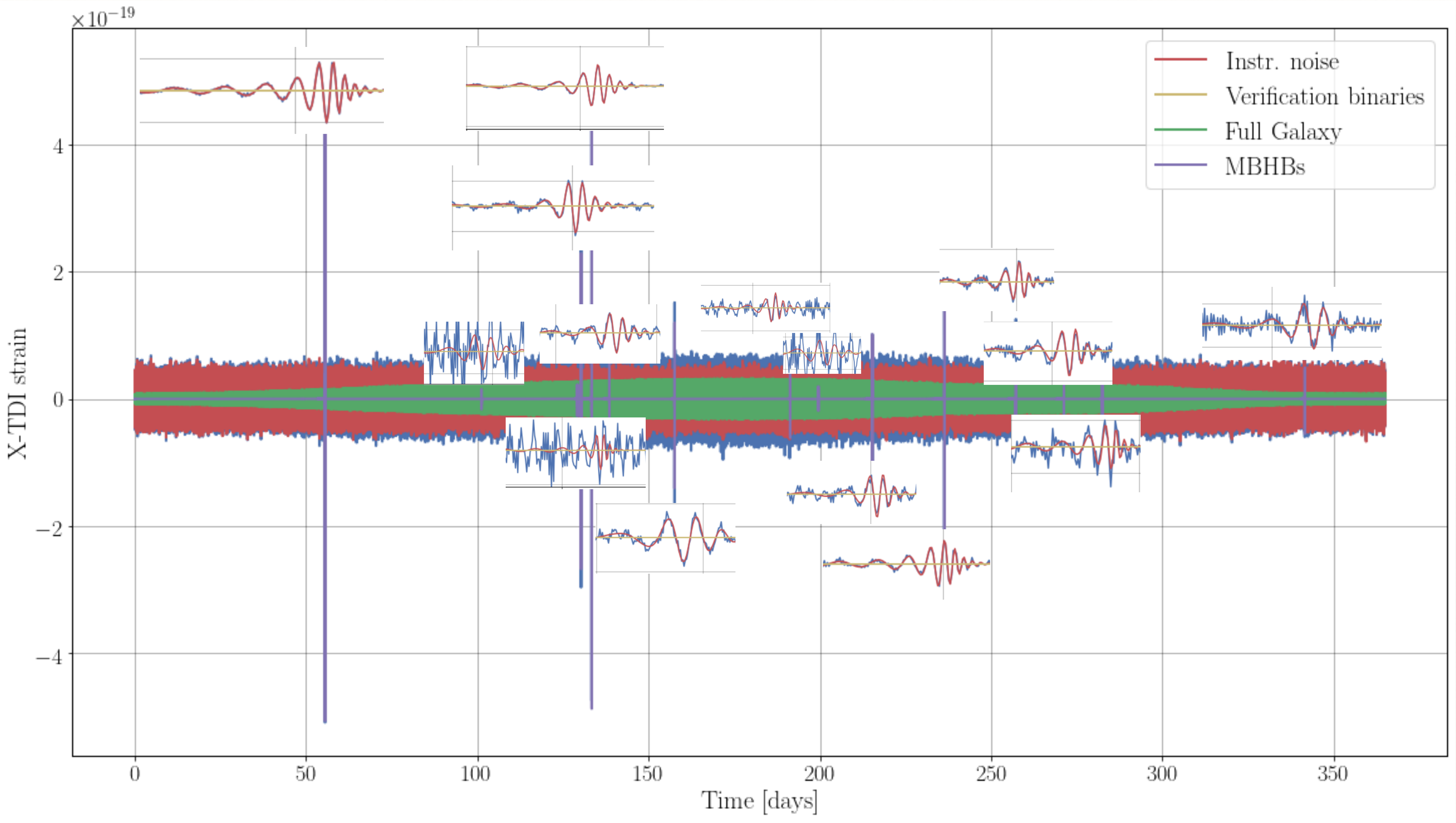


“Sangria” in time domain



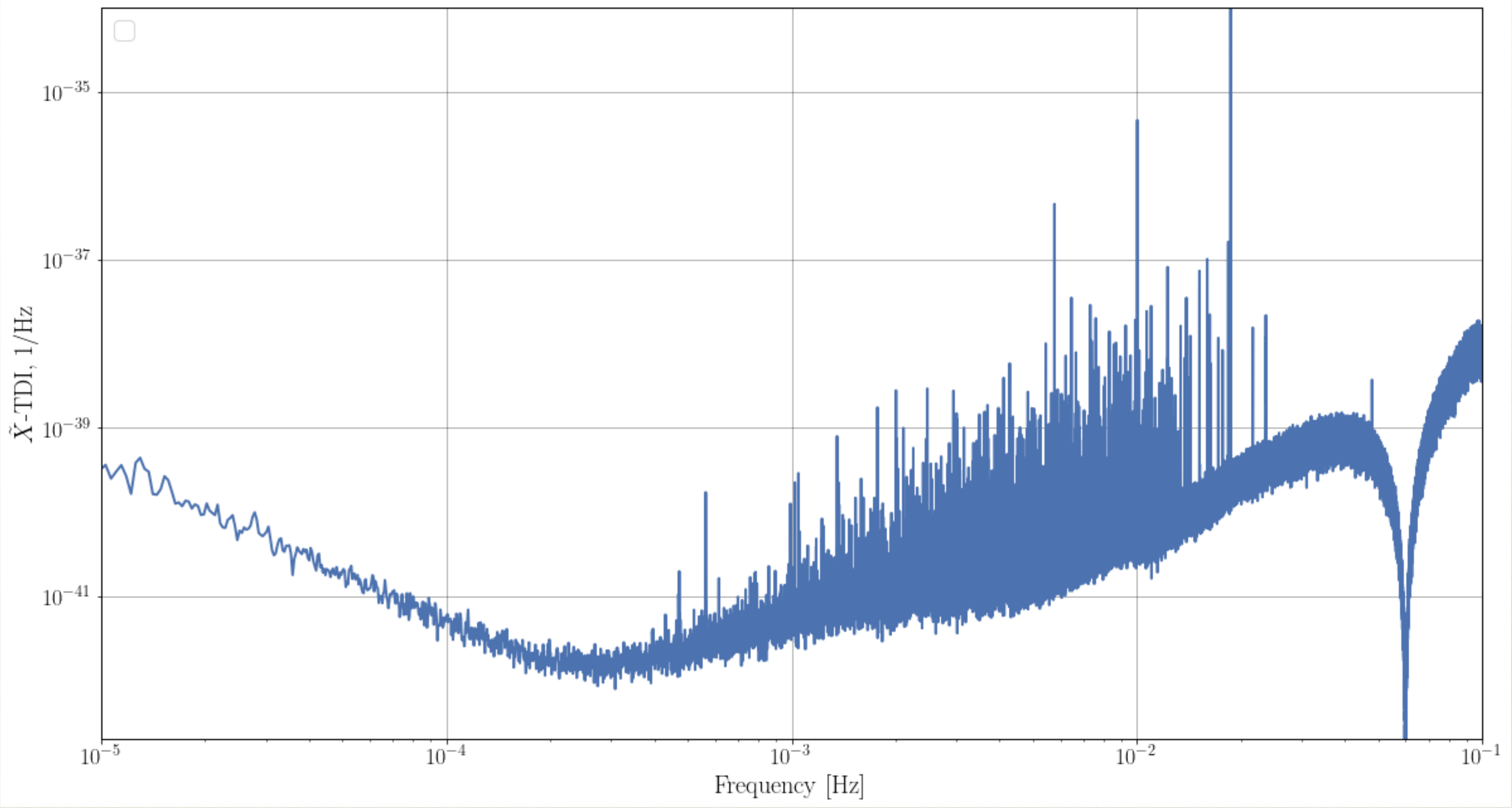
lisa

“Sangria” in time domain



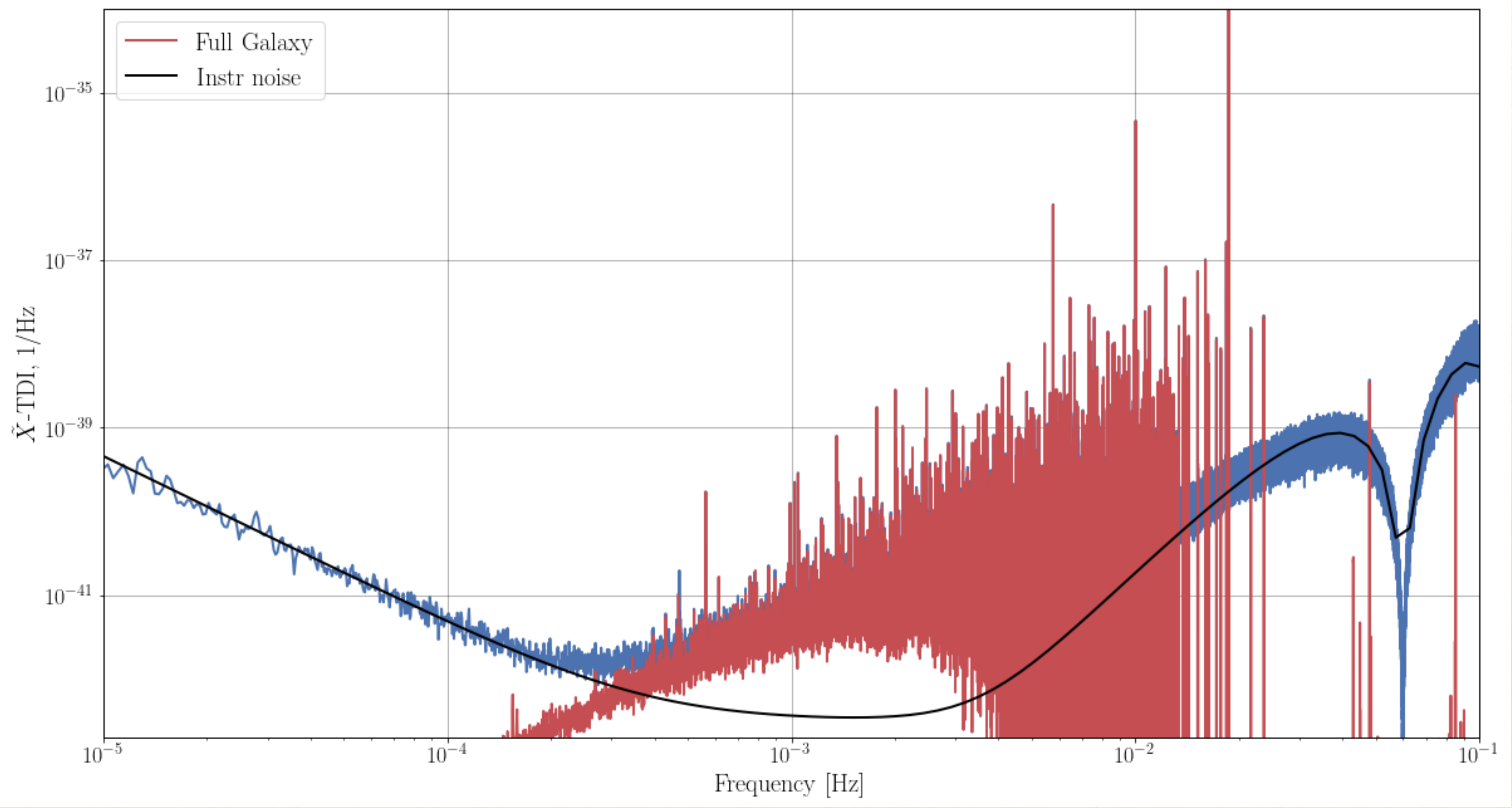
lisa

“Sangria” in frequency domain



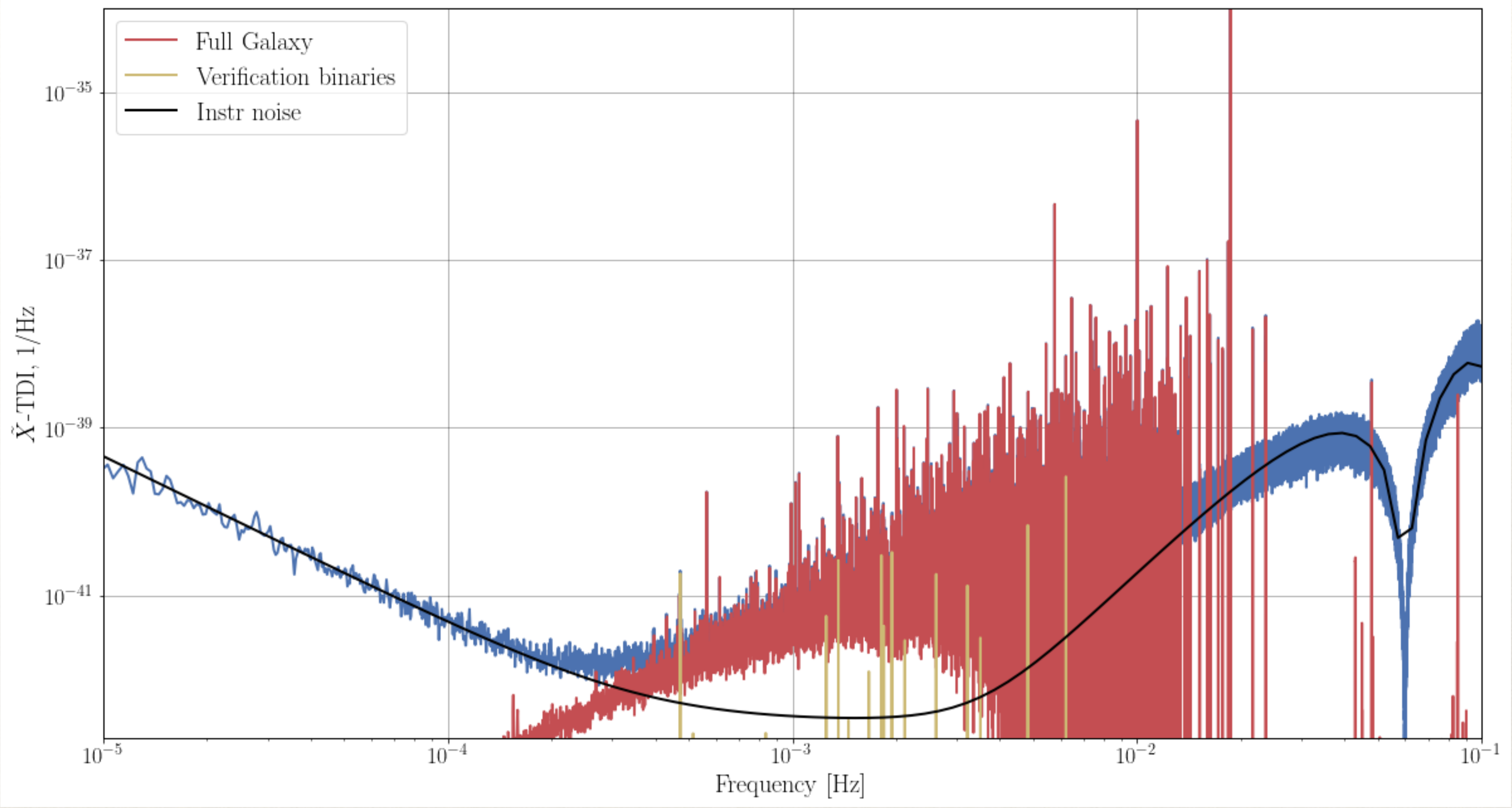
lisa

“Sangria” in frequency domain



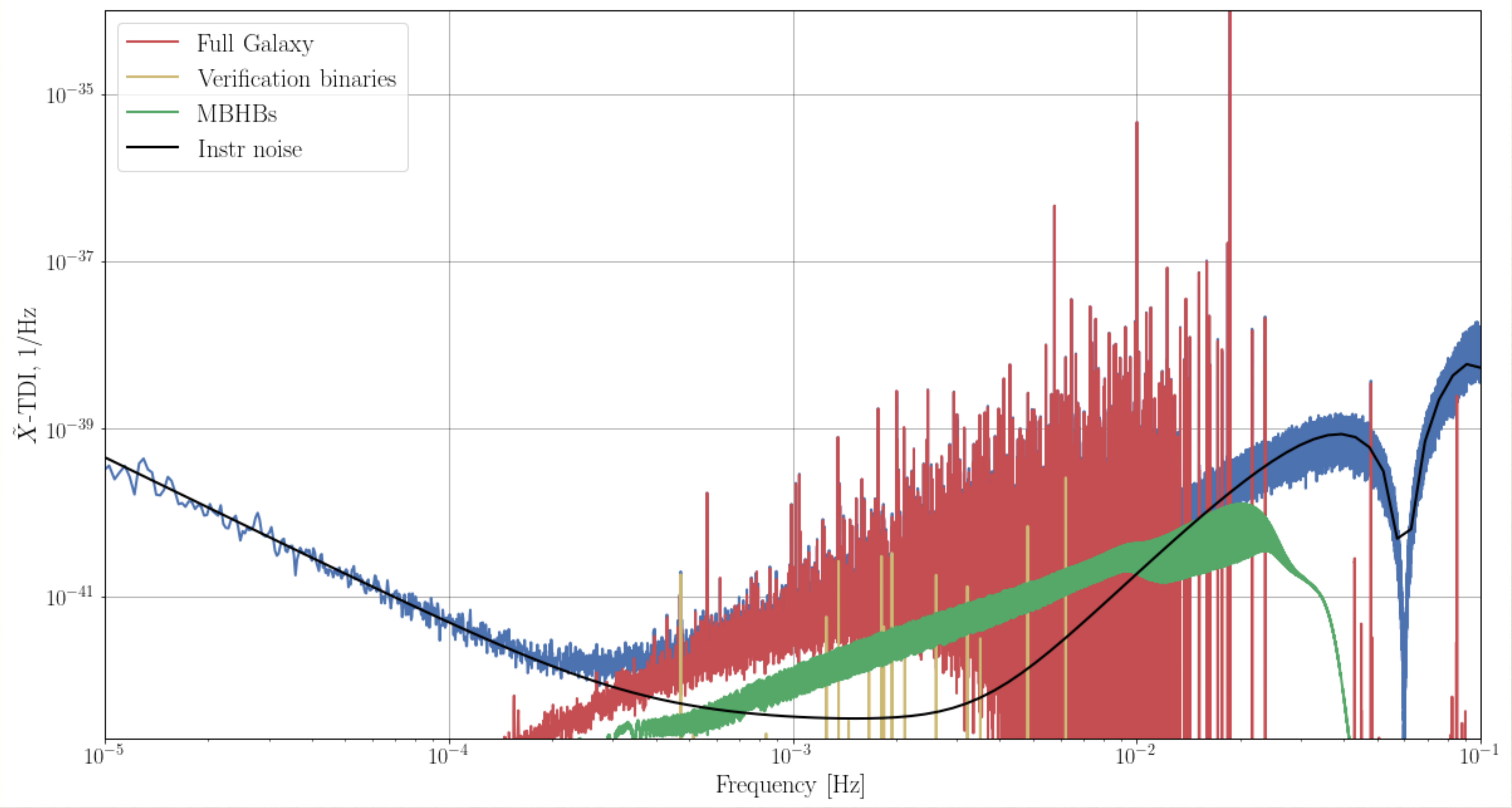
lisa

“Sangria” in frequency domain



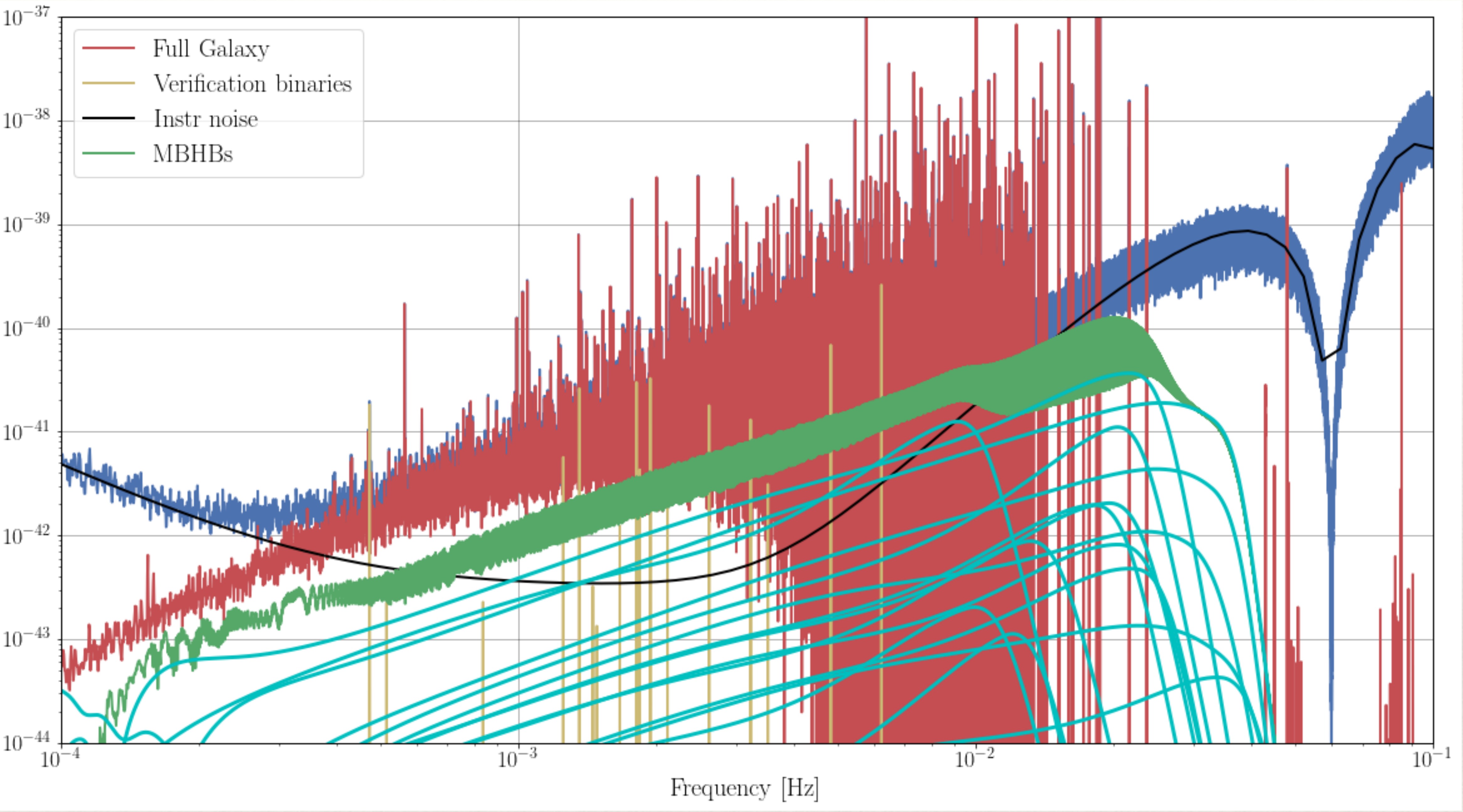
lisa

“Sangria” in frequency domain



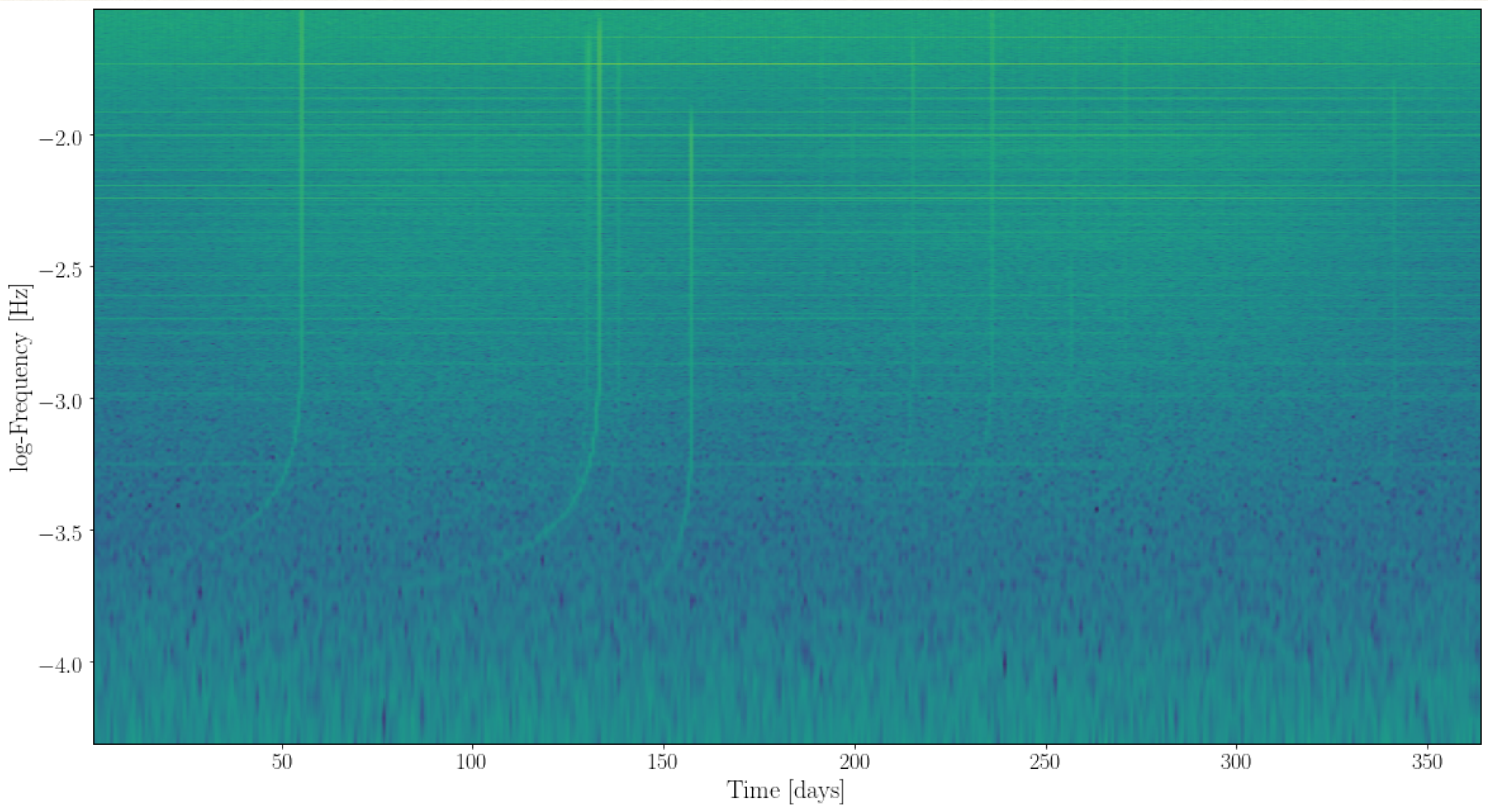
lisa

“Sangria” in frequency domain



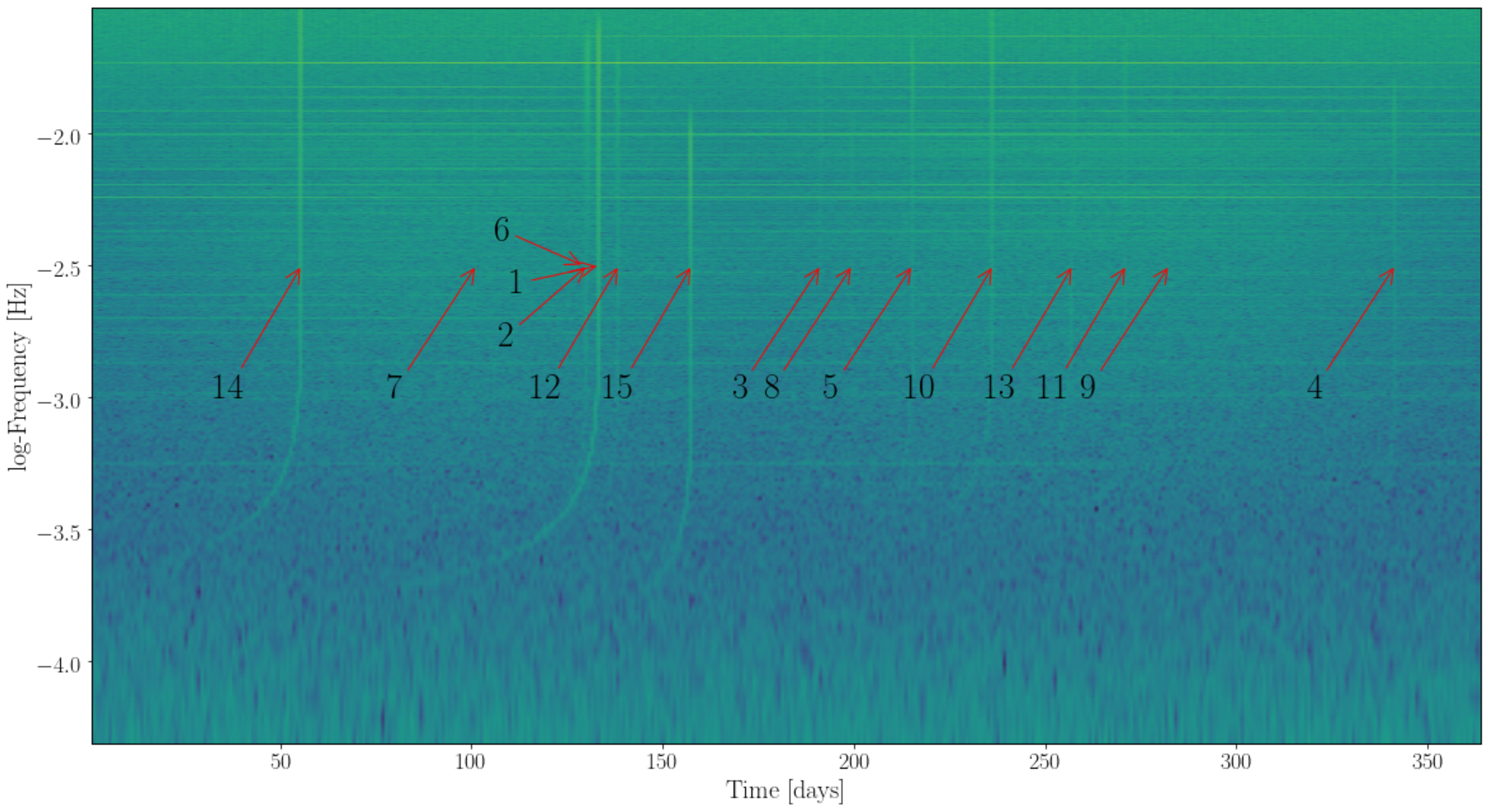
lisa

“Sangria” in time-frequency



lisa

“Sangria” in time-frequency



lisa

LISA data challenge-2b (“Spritz”)

- Spritz (alcoholic beverage), an aperitif consisting of wine, sparkling water, and liqueur

- Two data sets:
  - MBHB: 3 month of data
  - Verification binaries, 1 year long
- Both data sets will contain
  - scheduled gaps
  - glitches from LISAPathfinder
  - 2nd order Keplerian LISA orbit
  - Non-stationary noise: Galactic foreground

lisa

LISA data challenge-1b (“Yorsh”)

Yorsh (Also known as **Mora Grogg**) (Russian: Ёрш which means "Ruffe") is a Russian mixed drink consisting of beer thoroughly mixed with an ample quantity of vodka. It is traditionally drunk in a social setting, typically with a toast followed by downing a full glass of it at one go.

- Two types of GW sources (left over from challenge1b):
  - SBBH
    - 1 source with SNR ~ 25, 2 years
    - 1 source with SNR ~ 10, 2 years
  - EMRI
    - 1 source, SNR ~40, generic orbit, simple model (AAK)
    - 1 source, SNR ~40, non-spining MBH, fully relativistic

Conclusion

- LISA is not LIGO in space
- But… We can learn a lot from LVC and use methods/techniques in LISA data analysis.
- Many challenges caused by
  - multimodality of likelihood,
  - systematics in GW models,
  - large number of overlapped signals
  - non-stationarity of the noise (gaps, glitches, ….)
  - complex response,
  - strong/long duration of signals

- "Sangria" data set is served: download it and see what you we can extract.

lisa