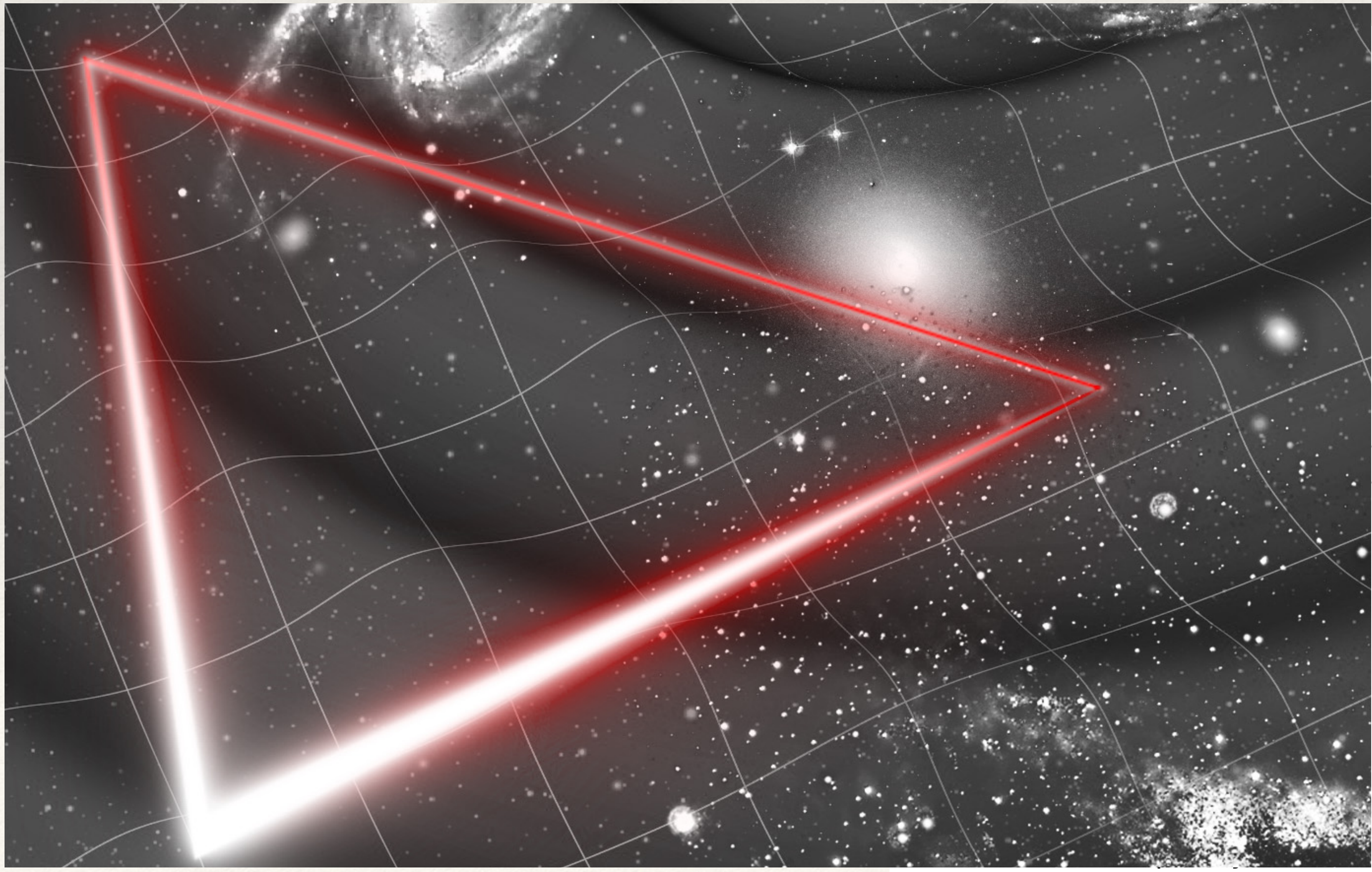
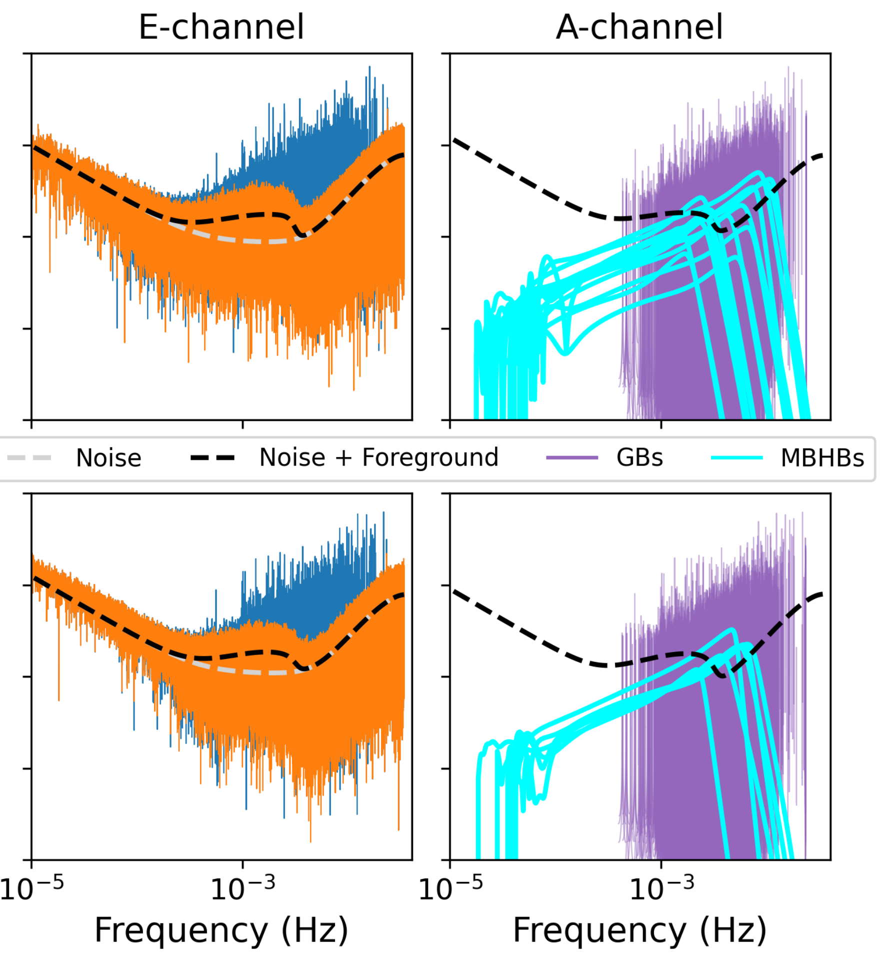
Gravitational waves coming at you from all directions: challenges in data analysis for LISA

Jonathan Gair, Max Planck Institute for Gravitational Physics (Potsdam)

Heterogeneous Data and Large Representation Models in Science,

Toulouse, October 1st 2024

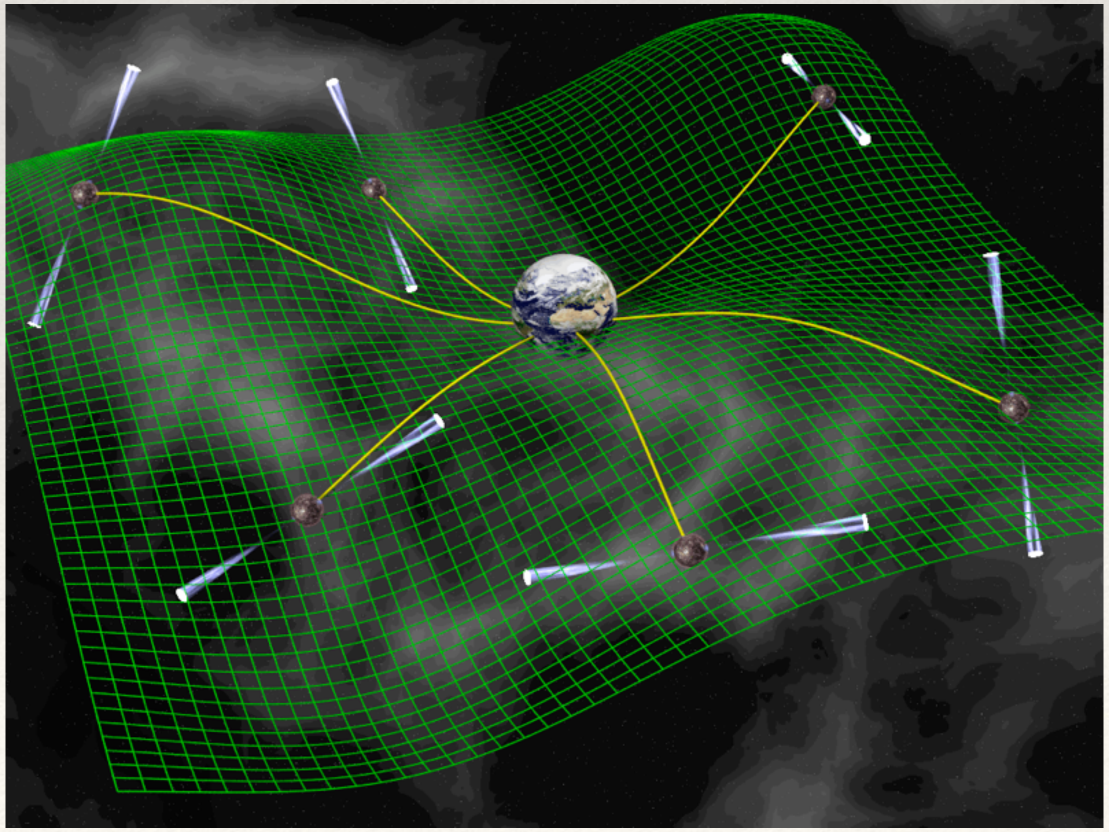




Talk outline

- Context: current gravitational wave detectors
- The Laser Interferometer Space Antenna
- Current state of the art in LISA data analysis development
- Key outstanding challenges in LISA data analysis
- New approaches

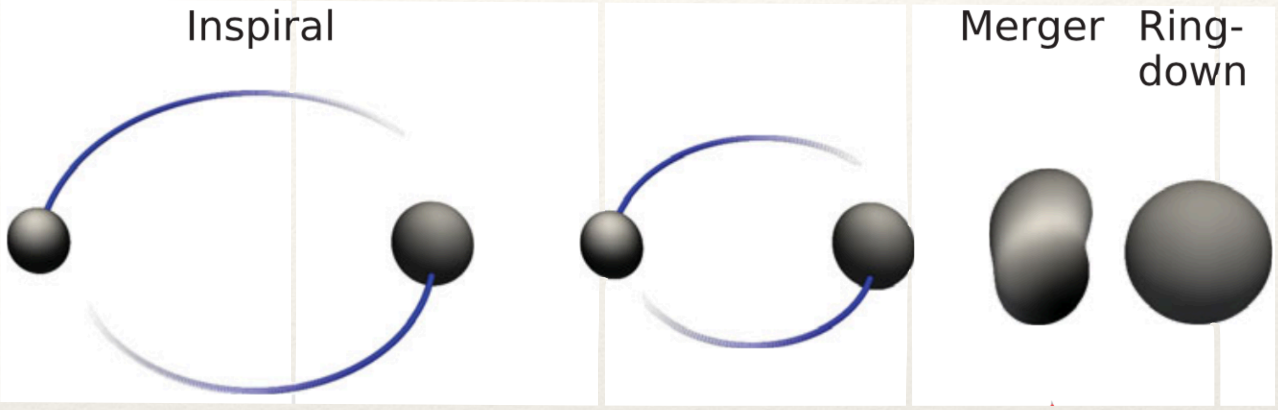
Context: gravitational wave detectors

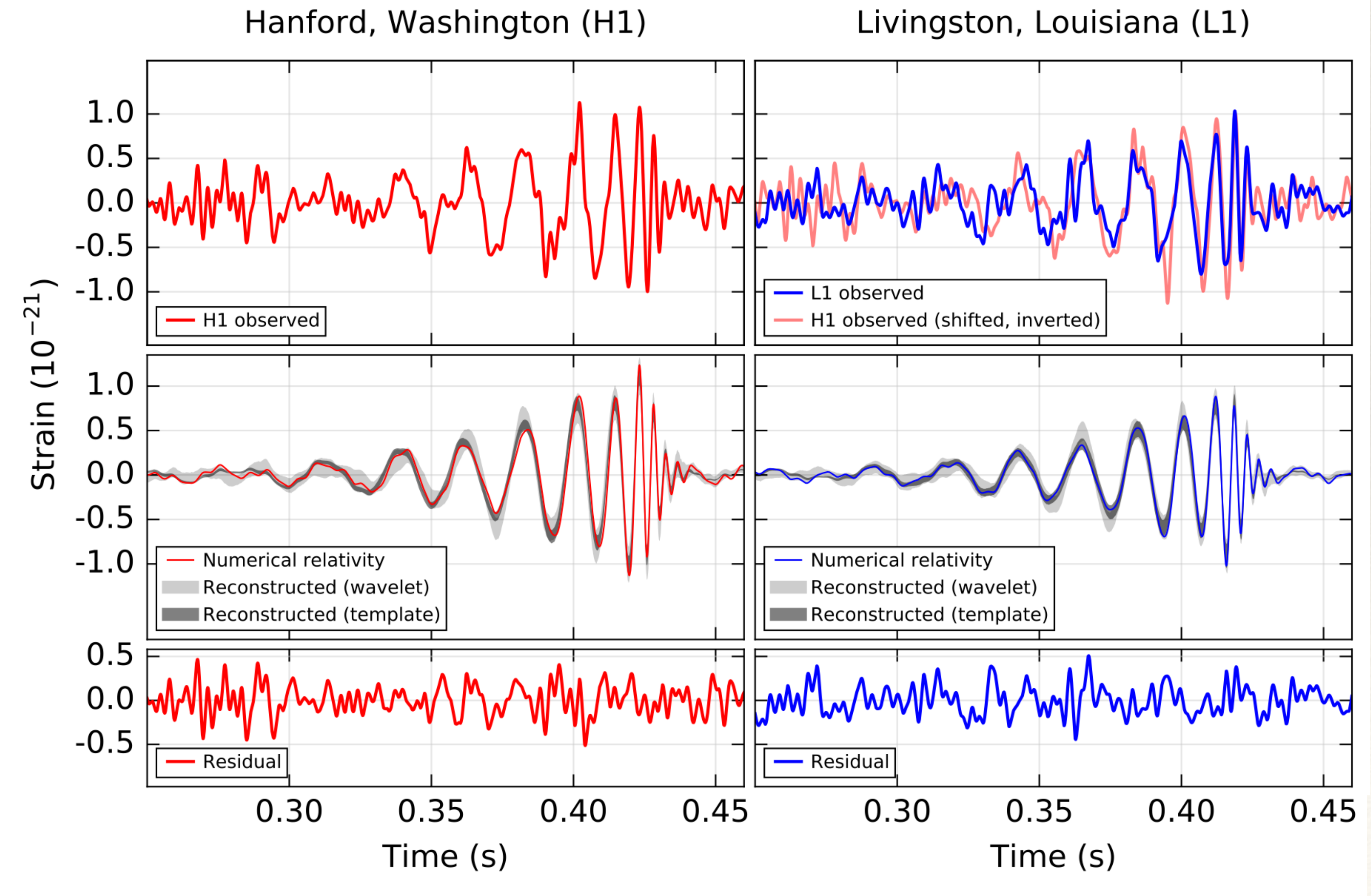


- A network of ***ground-based gravitational wave interferometers*** is currently operating
    - ***LIGO***: two 4km interferometers in WA and LA. Operating since September 2015.
    - ***Virgo***: 3km interferometer near Pisa, Italy (since 2017). ***KAGRA***: Japanese 4km underground detector (since 2020).
- ***Pulsar timing arrays*** are searching for nanohertz gravitational waves by accurate timing of millisecond pulsars
    - Several major collaborations, including *NANOGrav*, *PPTA*, *CPTA* and the ***EPTA***.

Context: first detection

- Merging Binary Black Hole, GW150914, at a distance of ~400 Mpc.
- Masses: $29M_{\odot} + 36M_{\odot} \rightarrow 62M_{\odot}$
- Signal fully consistent with General Relativity.





Nobel prize 2017

Photo: Bryce Vickmark
Rainer Weiss

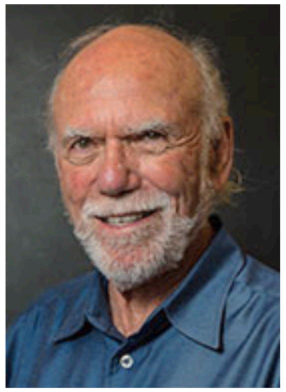
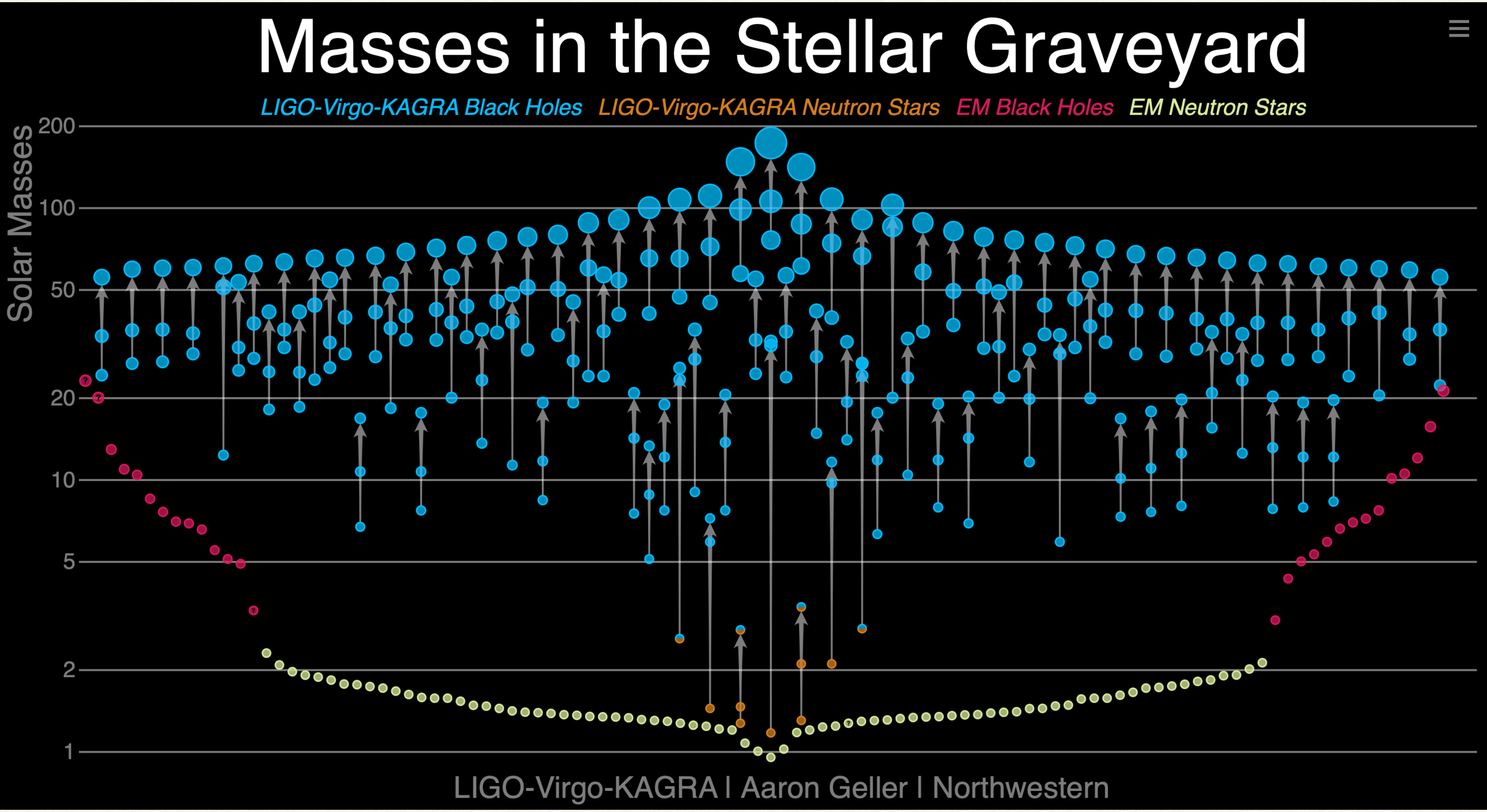


Photo: Caltech
Barry C. Barish

Photo: Caltech Alumni Association
Kip S. Thorne

Context: LIGO/Virgo observations



Context: PTA observations

- In June 2023, the major PTAs announced a likely detection of a GW background.
- Key signature is a characteristic correlation pattern between pulsars in different sky locations.
- Current data supports this correlation at ~2-4σ.

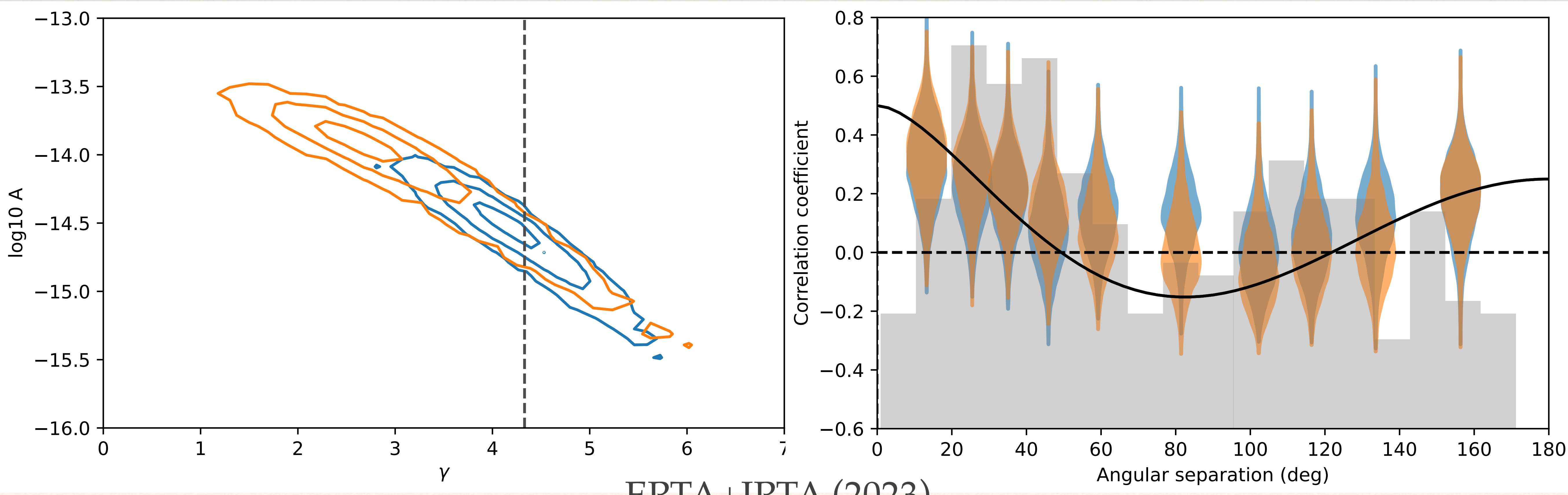
ZEIT ONLINE

Astrophysik

Neue Signale aus den Tiefen des Universums

Dem Verständnis des Alls etwas näher: Forscherteams weltweit könnten erstmals Gravitationswellen gigantischer Schwarzer Löcher entdeckt haben

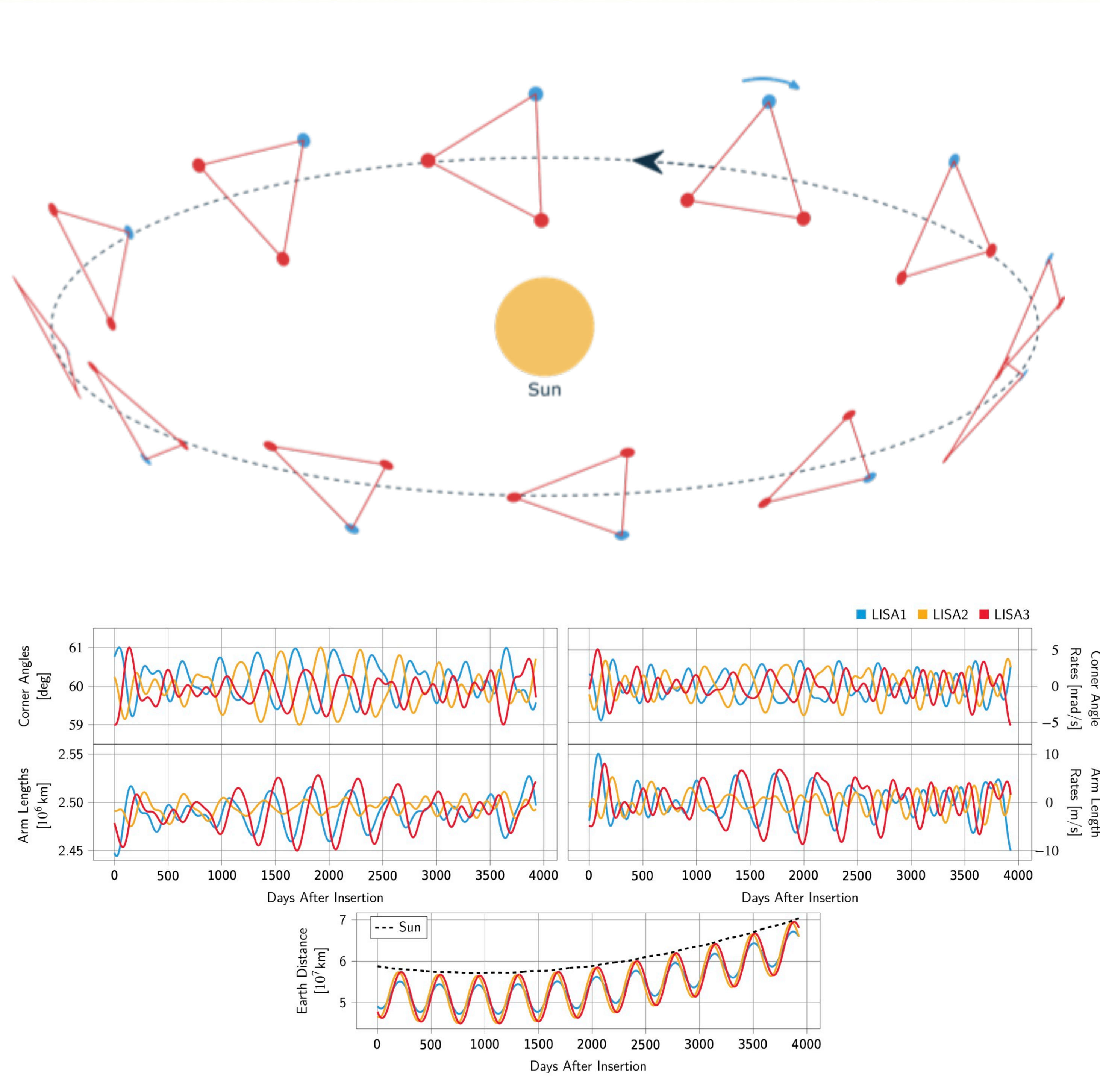
Von **Viola Kiel**



EPTA+IPTA (2023)

The Laser Interferometer Space Antenna

- LISA will comprise three satellites, ~2.5km apart (± 2%), in a heliocentric, earth-trailing orbit.
- Two laser links (one in each direction) connecting each pair of satellites.
- Constellation between 50 and 70 million km from Earth in first ten years – gradually drifts away.
- ESA-led, but NASA is a significant junior partner.
- Technology demonstrator mission, LISA Pathfinder, launched 2015.



The Laser Interferometer Space Antenna

- LISA was officially adopted as an ESA mission at the SPC meeting on January 25th 2024. Launch date: second half of 2035.

→ THE EUROPEAN SPACE AGENCY

SCIENCE & EXPLORATION

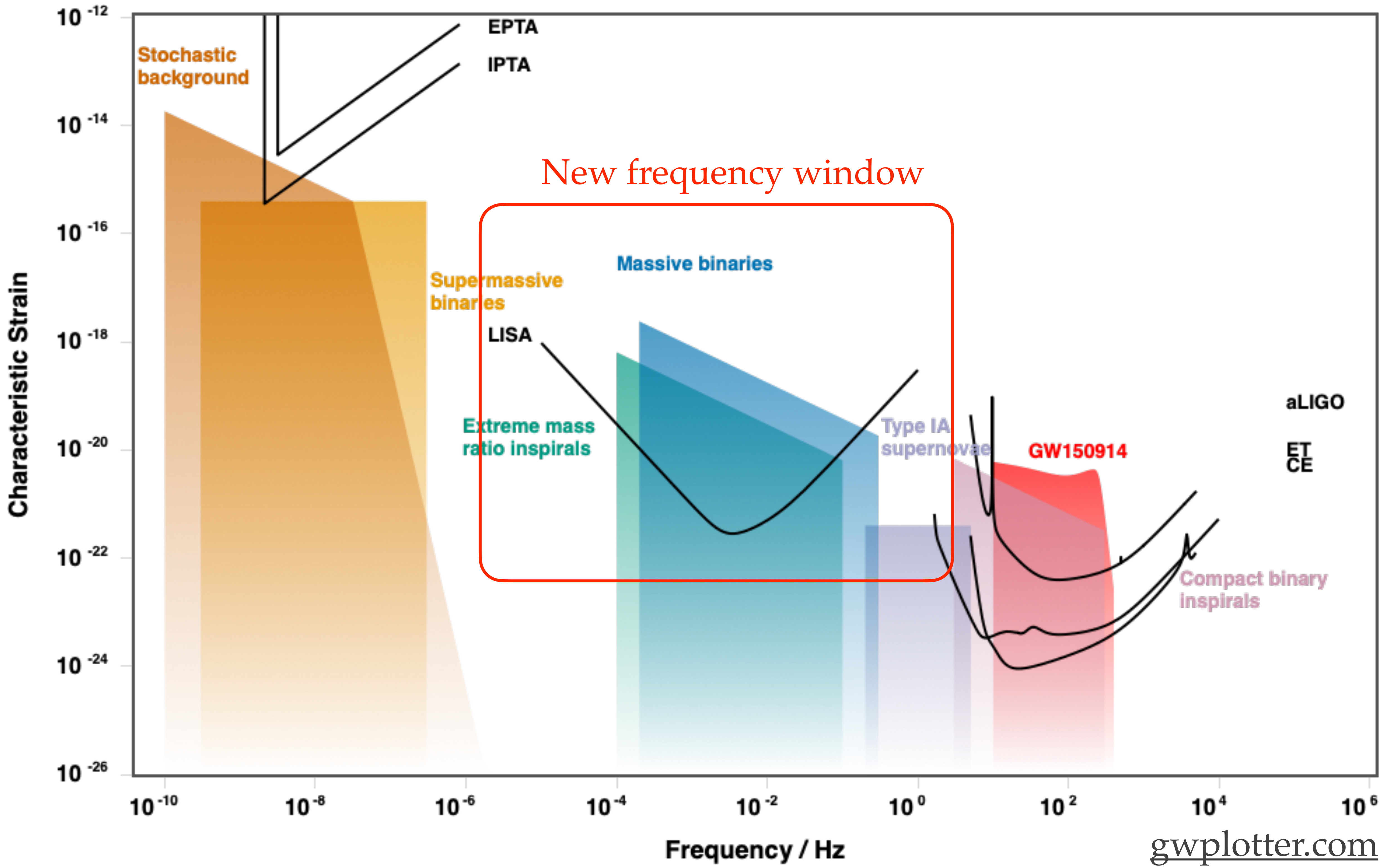
Capturing the ripples of spacetime: LISA gets go-ahead

25/01/2024 41968 VIEWS 201 LIKES

ESA / Science & Exploration / Space Science

Today, ESA's Science Programme Committee approved the Laser Interferometer Space Antenna (LISA) mission, the first scientific endeavour to detect and study gravitational waves from space.

GW frequency spectrum



gwplotter.com

GW sources for LISA (heterogeneous data)

- LISA is expected to observe gravitational waves from
  - ***Ultra-compact binaries (UCBs)***: binaries of stellar compact objects in the Milky Way with ~hour long periods. Dominated by double white dwarf binaries. Total population of **~10^7 systems**, of which **~10^4 resolvable** and the rest form a foreground. Signals essentially monochromatic and last **entire duration of mission**.

GW sources for LISA (heterogeneous data)

- LISA is expected to observe gravitational waves from
  - ***Ultra-compact binaries (UCBs)***: binaries of stellar compact objects in the Milky Way with ~hour long periods. Dominated by double white dwarf binaries. Total population of **~10^7 systems**, of which **~10^4 resolvable** and the rest form a foreground. Signals essentially monochromatic and last **entire duration of mission**.
  - ***Massive black hole binaries (MBHBs)***: binaries of black holes with mass ~10^4 — 10^7 solar masses form following mergers of their host galaxies. In band for up to **a few months** and very loud. Rate uncertain, but could be **several tens per year**.

GW sources for LISA (heterogeneous data)

- LISA is expected to observe gravitational waves from
  - ***Ultra-compact binaries (UCBs)***: binaries of stellar compact objects in the Milky Way with ~hour long periods. Dominated by double white dwarf binaries. Total population of **~10^7 systems**, of which **~10^4 resolvable** and the rest form a foreground. Signals essentially monochromatic and last **entire duration of mission**.
  - ***Massive black hole binaries (MBHBs)***: binaries of black holes with mass ~10^4 — 10^7 solar masses form following mergers of their host galaxies. In band for up to **a few months** and very loud. Rate uncertain, but could be **several tens per year**.
  - ***Extreme-mass-ratio inspirals (EMRIs)***: mergers of compact objects formed from stars with massive black holes in the centres of galaxies. **In band for O(yrs)** before merger. **Rate** unknown, could be as low as ~1 yr^{-1} or **as high as 1000 yr^{-1}**.

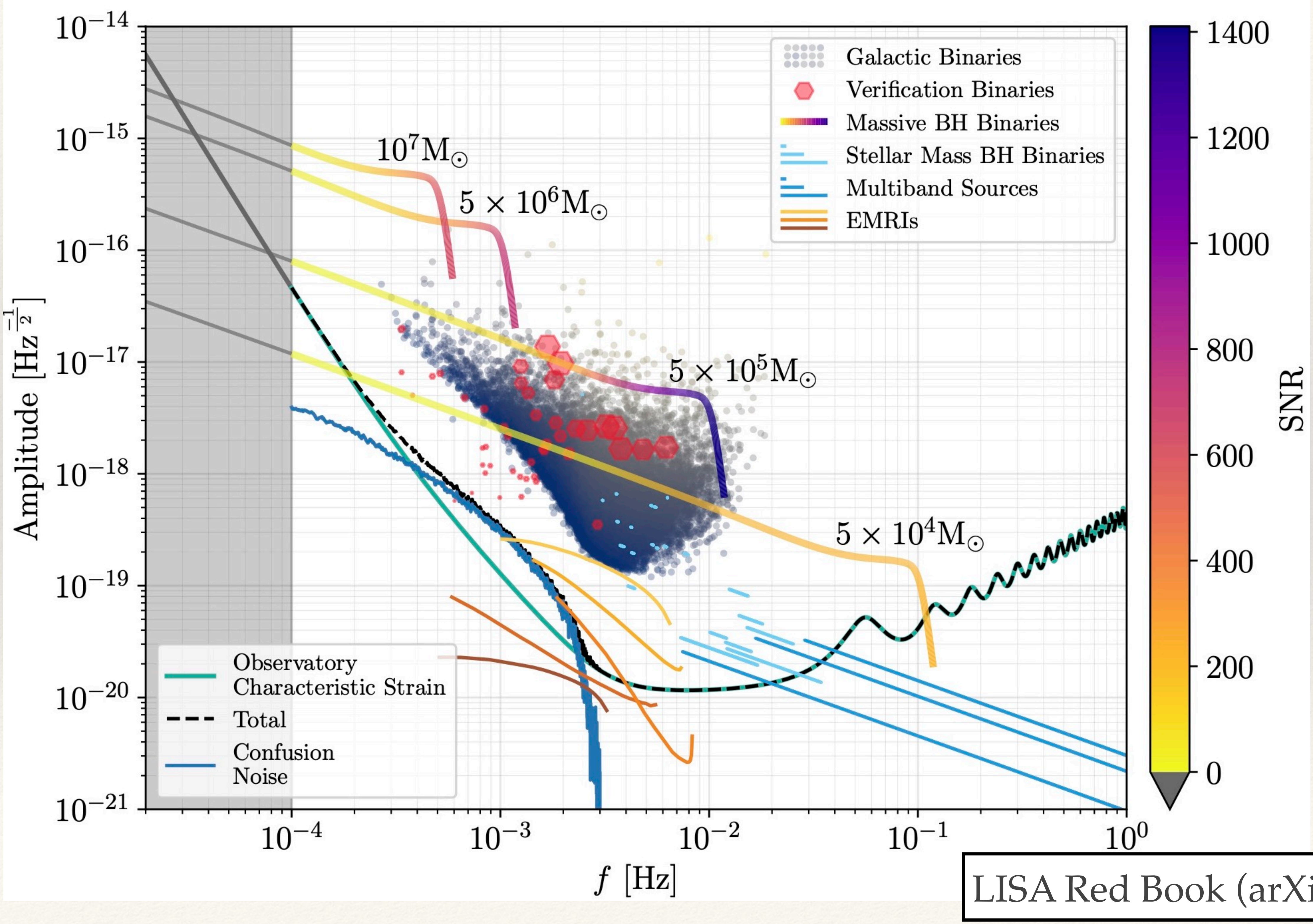
GW sources for LISA (heterogeneous data)

- LISA is expected to observe gravitational waves from
    - ***Ultra-compact binaries (UCBs)***: binaries of stellar compact objects in the Milky Way with ~hour long periods. Dominated by double white dwarf binaries. Total population of **~10^7 systems**, of which **~10^4 resolvable** and the rest form a foreground. Signals essentially monochromatic and last **entire duration of mission**.
    - ***Massive black hole binaries (MBHBs)***: binaries of black holes with mass ~10^4 — 10^7 solar masses form following mergers of their host galaxies. In band for up to **a few months** and very loud. Rate uncertain, but could be **several tens per year**.
    - ***Extreme-mass-ratio inspirals (EMRIs)***: mergers of compact objects formed from stars with massive black holes in the centres of galaxies. **In band for O(yrs)** before merger. **Rate** unknown, could be as low as ~1 yr^{-1} or **as high as 1000 yr^{-1}**.
    - ***Stellar-origin-black-hole inspirals (SOBH)***: most massive binaries observed by LIGO/Virgo would be visible to LISA a few years before merger. **O(few) yr^{-1}**.

GW sources for LISA (heterogeneous data)

- LISA is expected to observe gravitational waves from
  - ***Ultra-compact binaries (UCBs)***: binaries of stellar compact objects in the Milky Way with ~hour long periods. Dominated by double white dwarf binaries. Total population of ~**10^7 systems**, of which ~**10^4 resolvable** and the rest form a foreground. Signals essentially monochromatic and last **entire duration of mission**.
  - ***Massive black hole binaries (MBHBs)***: binaries of black holes with mass ~10^4 — 10^7 solar masses form following mergers of their host galaxies. In band for up to **a few months** and very loud. Rate uncertain, but could be **several tens per year**.
  - ***Extreme-mass-ratio inspirals (EMRIs)***: mergers of compact objects formed from stars with massive black holes in the centres of galaxies. **In band for O(yrs)** before merger. **Rate** unknown, could be as low as ~1 yr^{-1} or **as high as 1000 yr^{-1}**.
  - ***Stellar-origin-black-hole inspirals (SOBH)***: most massive binaries observed by LIGO/Virgo would be visible to LISA a few years before merger. **O(few) yr^{-1}**.
  - ***Cosmological sources***: phase transitions and other processes in the early Universe can generate **stochastic backgrounds** at mHz frequencies. Could also see individual bursts or a background generated by cosmic strings. **Amplitude/rate very uncertain**.

LISA data complexity



LISA Red Book (arXiv:2402.07571)

Key differences: LISA vs LIGO/Virgo

- There are similarities between data analysis for LISA and ground-based detectors (non-pointable detectors, signals buried in noise), but also several key differences
    - ***Signal duration***: primary source for LIGO/Virgo are compact binary mergers, which last ~O(1s) for BBHs, and up to O(1m) for BNS. LISA sources last between days (heavy MBHBs) to years (EMRIs) to entire mission (UCBs).

Key differences: LISA vs LIGO/Virgo

- There are similarities between data analysis for LISA and ground-based detectors (non-pointable detectors, signals buried in noise), but also several key differences
    - ***Signal duration***: primary source for LIGO/Virgo are compact binary mergers, which last ~O(1s) for BBHs, and up to O(1m) for BNS. LISA sources last between days (heavy MBHBs) to years (EMRIs) to entire mission (UCBs).
    - ***Source density***: LIGO/Virgo observe sources at a rate of ~100/year. LISA data will contain 10^7 UCBs, O(10^3) EMRIs and O(10^2) MBHBs.

Key differences: LISA vs LIGO/Virgo

- There are similarities between data analysis for LISA and ground-based detectors (non-pointable detectors, signals buried in noise), but also several key differences
  - ***Signal duration***: primary source for LIGO/Virgo are compact binary mergers, which last ~O(1s) for BBHs, and up to O(1m) for BNS. LISA sources last between days (heavy MBHBs) to years (EMRIs) to entire mission (UCBs).
  - ***Source density***: LIGO/Virgo observe sources at a rate of ~100/year. LISA data will contain 10^7 UCBs, O(10^3) EMRIs and O(10^2) MBHBs.
  - ***Signal identification***: LIGO/Virgo separate *search* and *characterisation*. First step identifies interesting data segments. Second stage is parameter estimation. Cannot separate these steps in LISA as all data contains signals. Need *global fit*.

Key differences: LISA vs LIGO/Virgo

- There are similarities between data analysis for LISA and ground-based detectors (non-pointable detectors, signals buried in noise), but also several key differences
    - ***Signal duration***: primary source for LIGO/Virgo are compact binary mergers, which last ~O(1s) for BBHs, and up to O(1m) for BNS. LISA sources last between days (heavy MBHBs) to years (EMRIs) to entire mission (UCBs).
    - ***Source density***: LIGO/Virgo observe sources at a rate of ~100/year. LISA data will contain 10^7 UCBs, O(10^3) EMRIs and O(10^2) MBHBs.
    - ***Signal identification***: LIGO/Virgo separate *search* and *characterisation*. First step identifies interesting data segments. Second stage is parameter estimation. Cannot separate these steps in LISA as all data contains signals. Need *global fit*.
    - ***Number of independent detectors***: there are three independent ground-based detectors, with uncorrelated noise. LISA has two separate data channels, but not really independent. Requires simultaneous noise & signal estimation.

Key differences: LISA vs LIGO/Virgo

- There are similarities between data analysis for LISA and ground-based detectors (non-pointable detectors, signals buried in noise), but also several key differences
    - ***Signal duration***: primary source for LIGO/Virgo are compact binary mergers, which last ~O(1s) for BBHs, and up to O(1m) for BNS. LISA sources last between days (heavy MBHBs) to years (EMRIs) to entire mission (UCBs).
    - ***Source density***: LIGO/Virgo observe sources at a rate of ~100/year. LISA data will contain 10^7 UCBs, O(10^3) EMRIs and O(10^2) MBHBs.
    - ***Signal identification***: LIGO/Virgo separate *search* and *characterisation*. First step identifies interesting data segments. Second stage is parameter estimation. Cannot separate these steps in LISA as all data contains signals. Need *global fit*.
    - ***Number of independent detectors***: there are three independent ground-based detectors, with uncorrelated noise. LISA has two separate data channels, but not really independent. Requires simultaneous noise & signal estimation.
    - ***Instrumental artefacts***: data from both LIGO/Virgo and LISA contains glitches and data gaps, but these do not overlap most signals in LIGO/Virgo.

GW data analysis framework

- Data analysis for LISA is typically based on Bayesian inference methods applied to TDI time series data

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

- The noise is assumed to be Gaussian and stationary with a likelihood of the form

$$p(\mathbf{d}|\theta) \propto \exp\left[-\frac{1}{2}\left(\mathbf{d}-\mathbf{h}(\theta)|\mathbf{d}-\mathbf{h}(\theta)\right)\right] \qquad \mathbf{h}(\theta) = \sum_{i=1}^{n_t}\sum_{j=1}^{n_i} h_i(\theta_{i,j})$$

$$(a|b) = \int_{-\infty}^{\infty} \frac{\tilde{a}^*(f)\tilde{b}(f) + \tilde{a}(f)\tilde{b}^*(f)}{S_n(f)}\,\mathrm{d}f$$

- Complexity and computational cost of parameter estimation is driven by

GW data analysis framework

- Data analysis for LISA is typically based on Bayesian inference methods applied to TDI time series data

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

- The noise is assumed to be Gaussian and stationary with a likelihood of the form

$$p(\mathbf{d}|\theta) \propto \exp\left[-\frac{1}{2}\left(\mathbf{d}-\mathbf{h}(\theta)|\mathbf{d}-\mathbf{h}(\theta)\right)\right] \qquad \mathbf{h}(\theta) = \sum_{i=1}^{n_t}\sum_{j=1}^{n_i} h_i(\theta_{i,j})$$

$$(a|b) = \int_{-\infty}^{\infty} \frac{\tilde{a}^*(f)\tilde{b}(f) + \tilde{a}(f)\tilde{b}^*(f)}{S_n(f)}\,\mathrm{d}f$$

- Complexity and computational cost of parameter estimation is driven by
  - ***waveform model evaluation***: every likelihood evaluation requires the computation of many waveform models, which are expensive to evaluate.

GW data analysis framework

- Data analysis for LISA is typically based on Bayesian inference methods applied to TDI time series data

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

- The noise is assumed to be Gaussian and stationary with a likelihood of the form

$$p(\mathbf{d}|\theta) \propto \exp\left[-\frac{1}{2}\left(\mathbf{d}-\mathbf{h}(\theta)|\mathbf{d}-\mathbf{h}(\theta)\right)\right]$$

$$\mathbf{h}(\theta) = \sum_{i=1}^{n_t}\sum_{j=1}^{n_i} h_i(\theta_{i,j})$$

$$(a|b) = \int_{-\infty}^{\infty} \frac{\tilde{a}^*(f)\tilde{b}(f) + \tilde{a}(f)\tilde{b}^*(f)}{S_n(f)}\,\mathrm{d}f$$

- Complexity and computational cost of parameter estimation is driven by
  - ***waveform model evaluation***: every likelihood evaluation requires the computation of many waveform models, which are expensive to evaluate.
  - ***variable dimensionality***: number of sources of each type in data is unknown.

GW data analysis framework

- Data analysis for LISA is typically based on Bayesian inference methods applied to TDI time series data

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

- The noise is assumed to be Gaussian and stationary with a likelihood of the form

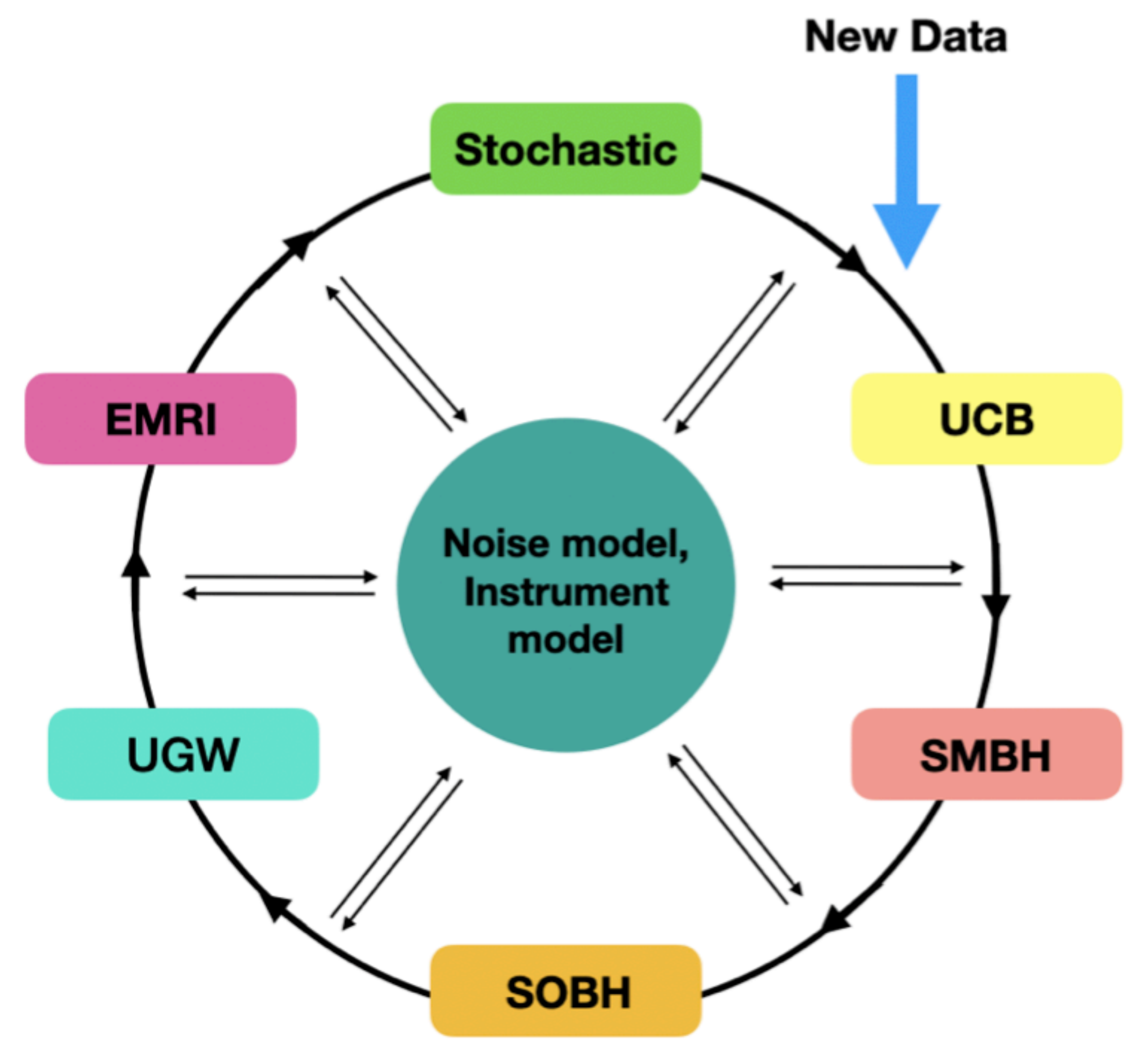
$$p(\mathbf{d}|\theta) \propto \exp\left[-\frac{1}{2}\left(\mathbf{d}-\mathbf{h}(\theta)|\mathbf{d}-\mathbf{h}(\theta)\right)\right] \qquad \mathbf{h}(\theta) = \sum_{i=1}^{n_t}\sum_{j=1}^{n_i} h_i(\theta_{i,j})$$

$$(a|b) = \int_{-\infty}^{\infty} \frac{\tilde{a}^*(f)\tilde{b}(f) + \tilde{a}(f)\tilde{b}^*(f)}{S_n(f)}\,\mathrm{d}f$$

- Complexity and computational cost of parameter estimation is driven by
    - ***waveform model evaluation***: every likelihood evaluation requires the computation of many waveform models, which are expensive to evaluate.
    - ***variable dimensionality***: number of sources of each type in data is unknown.
    - ***sampling***: we typically represent the complex posterior distribution by a set of samples. Drawing these requires ~millions of likelihood evaluation.

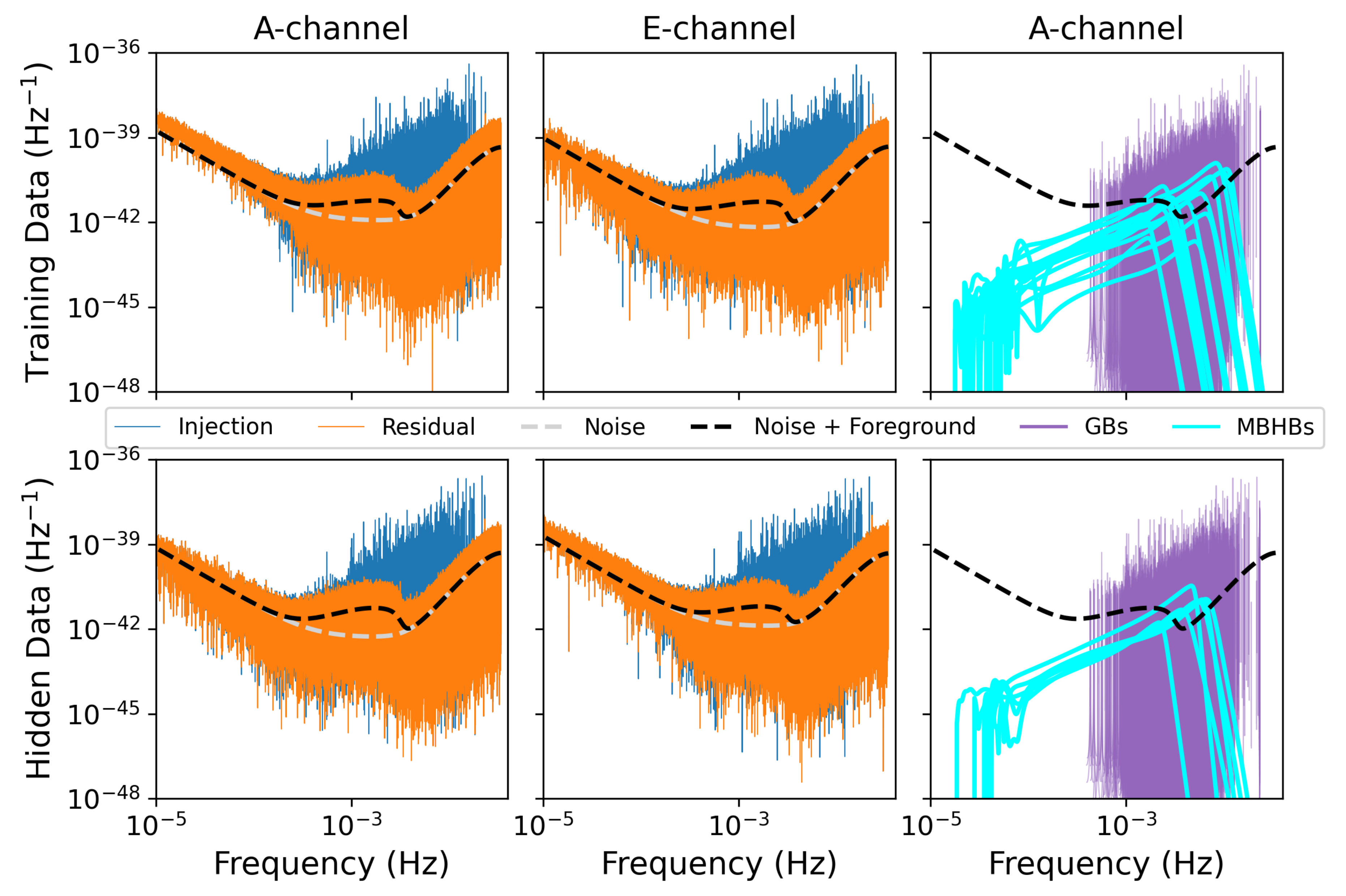
LISA Data Analysis

- LISA data set not big (few Gb) but the model is (large representation model).
- To date, successful solutions to the global fit problem have used classic stochastic sampling techniques.
- Typical strategy adopted is to iteratively update the solution for one source type and then move to the next. (Gibbs)
- Techniques like *reversible jump MCMC* are necessary to handle the problem of *variable dimensionality*.
- Employ *affine-invariant sampling* and *parallel tempering* to improve sampling convergence.



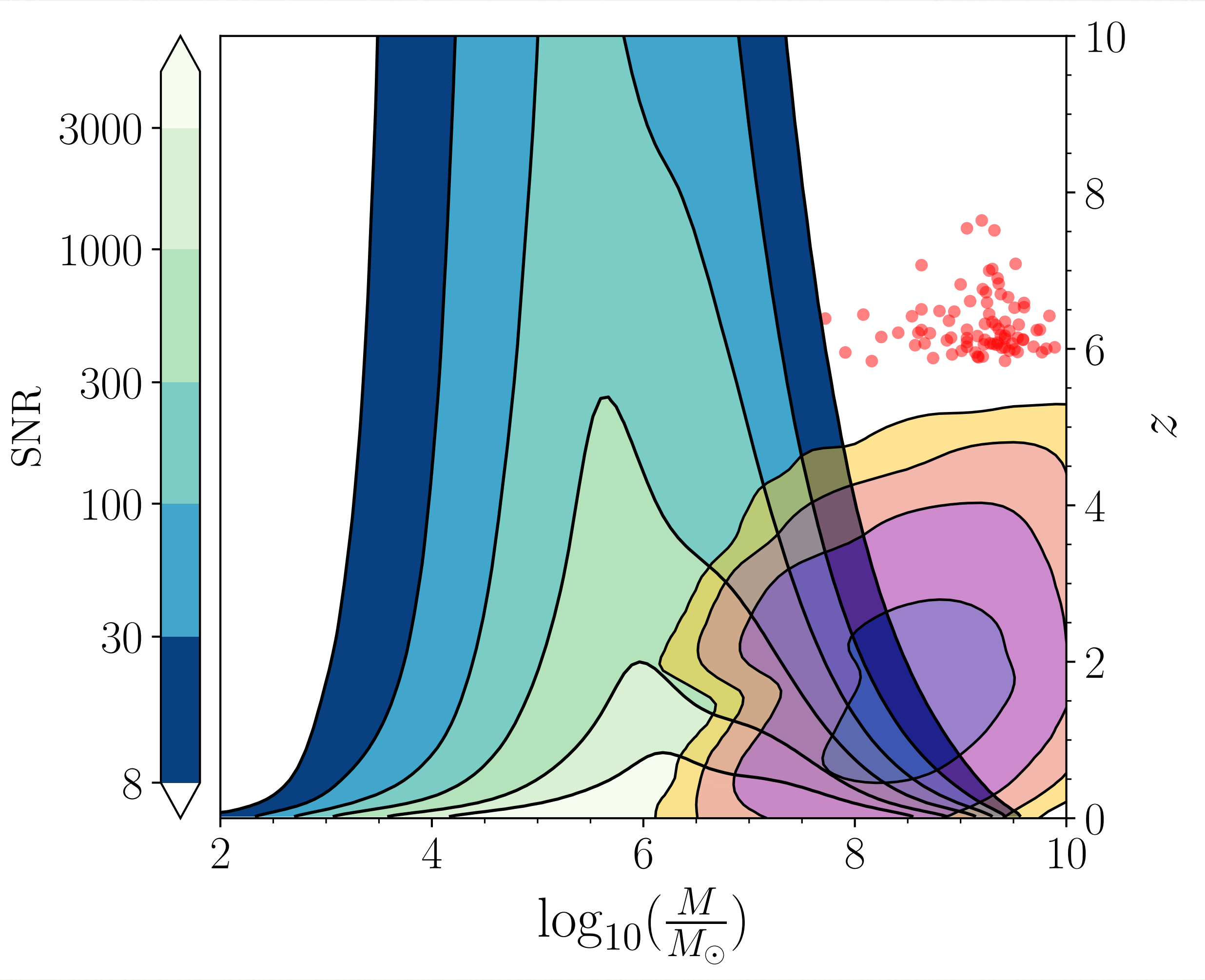
State of the art: *Sangria* data set

- LISA data analysis development is being promoted through a series of *Data Challenges*. Most sophisticated to date (Sangria): a galaxy of white dwarf binaries plus massive black hole signals in stationary Gaussian noise.



Strategy: massive black holes

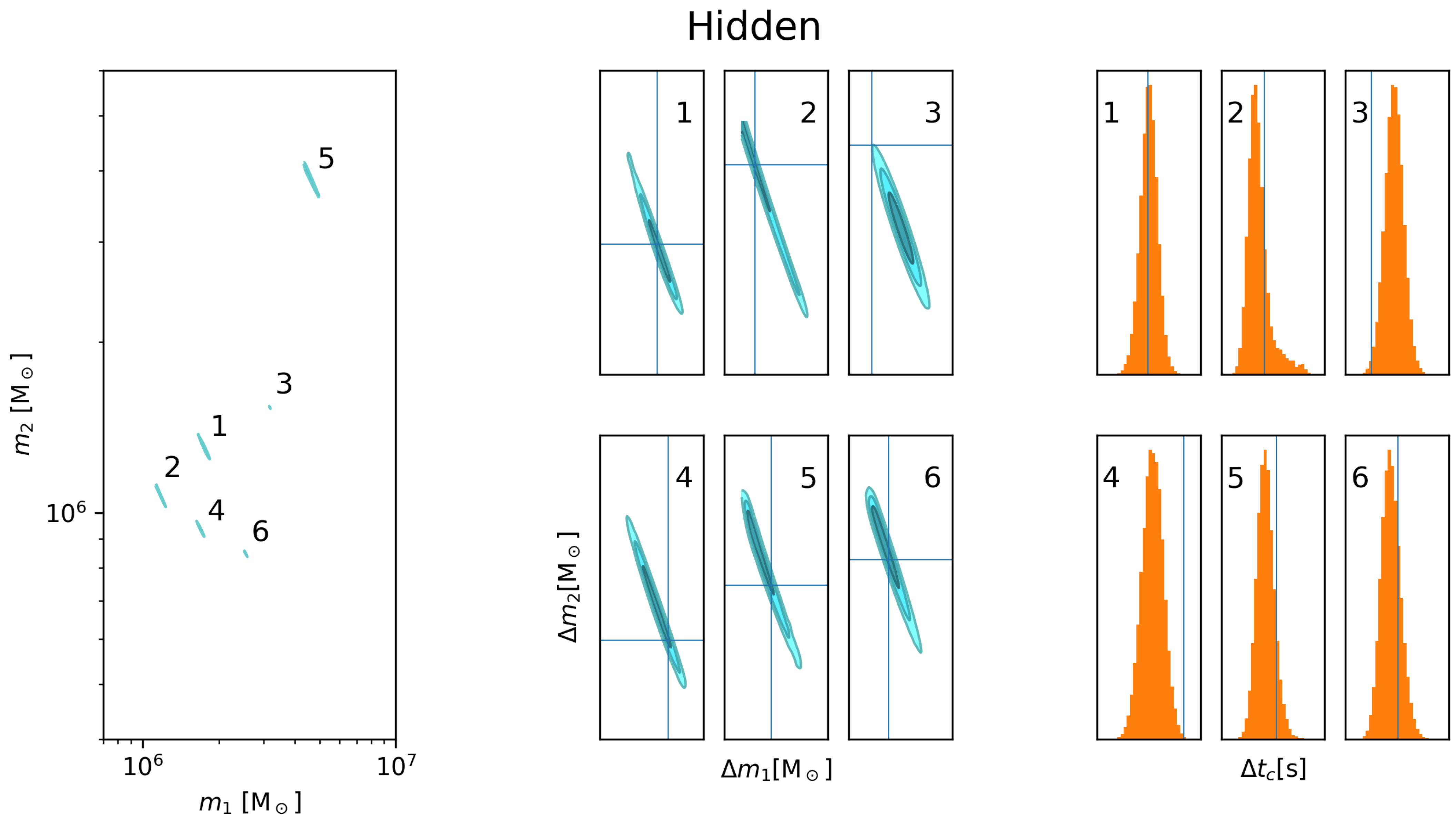
- Massive black holes can be observed with very high SNR by LISA. Merger typically stands out above the noise, so signals are *compact in time*.
- Data analysis uses
  - ***search phase***: sliding one day window used to identify mergers and crudely estimate parameters with stochastic search algorithms;
  - ***characterisation phase***: stochastic sampling of parameters, using initial estimates and *fixing number of sources*, used to obtain parameter posteriors.



LISA Red Book (arXiv:2402.07571)

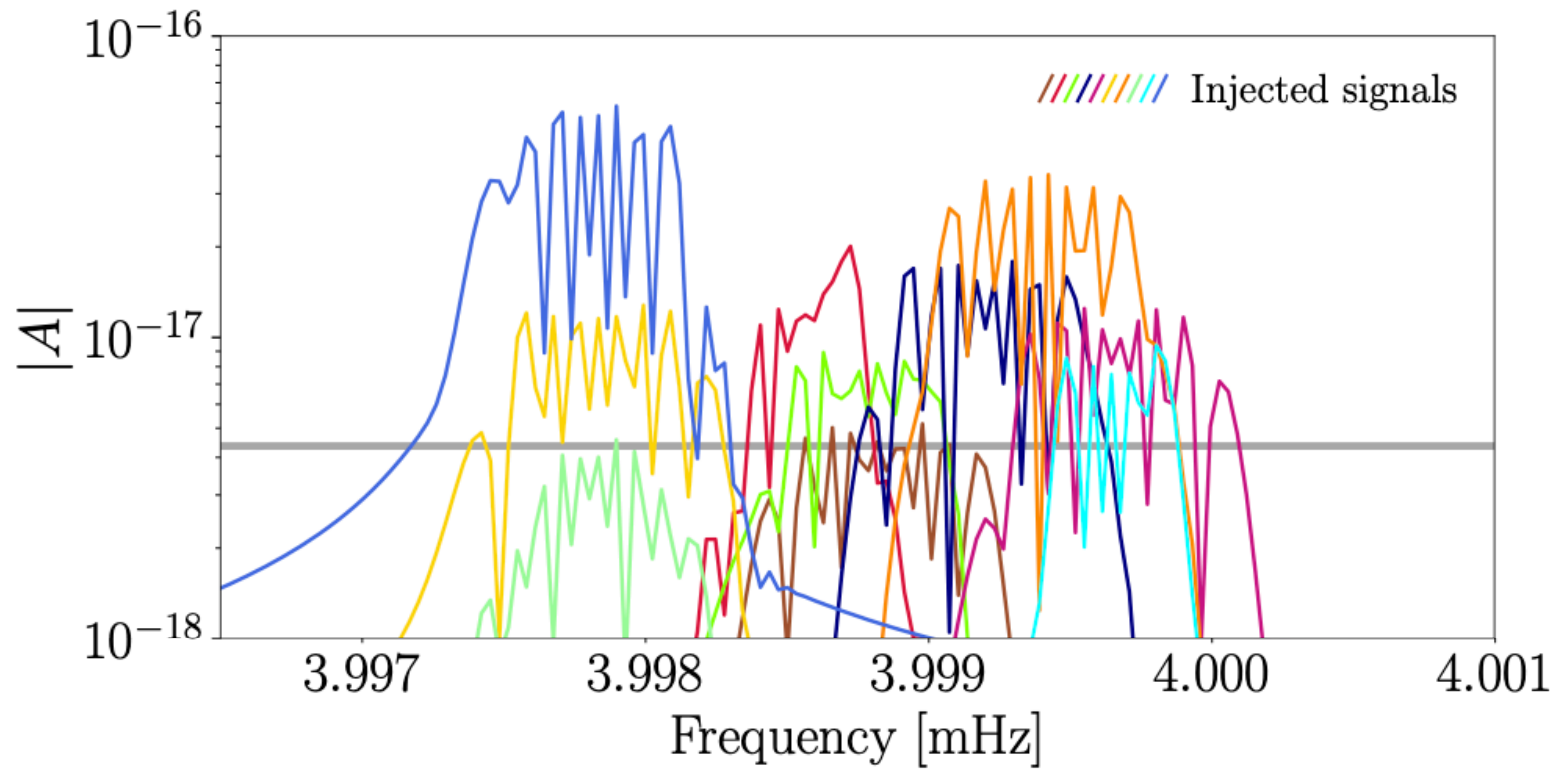
Results: massive black holes

- Massive black hole binary parameters determined to high precision and consistent with values used to generate data.



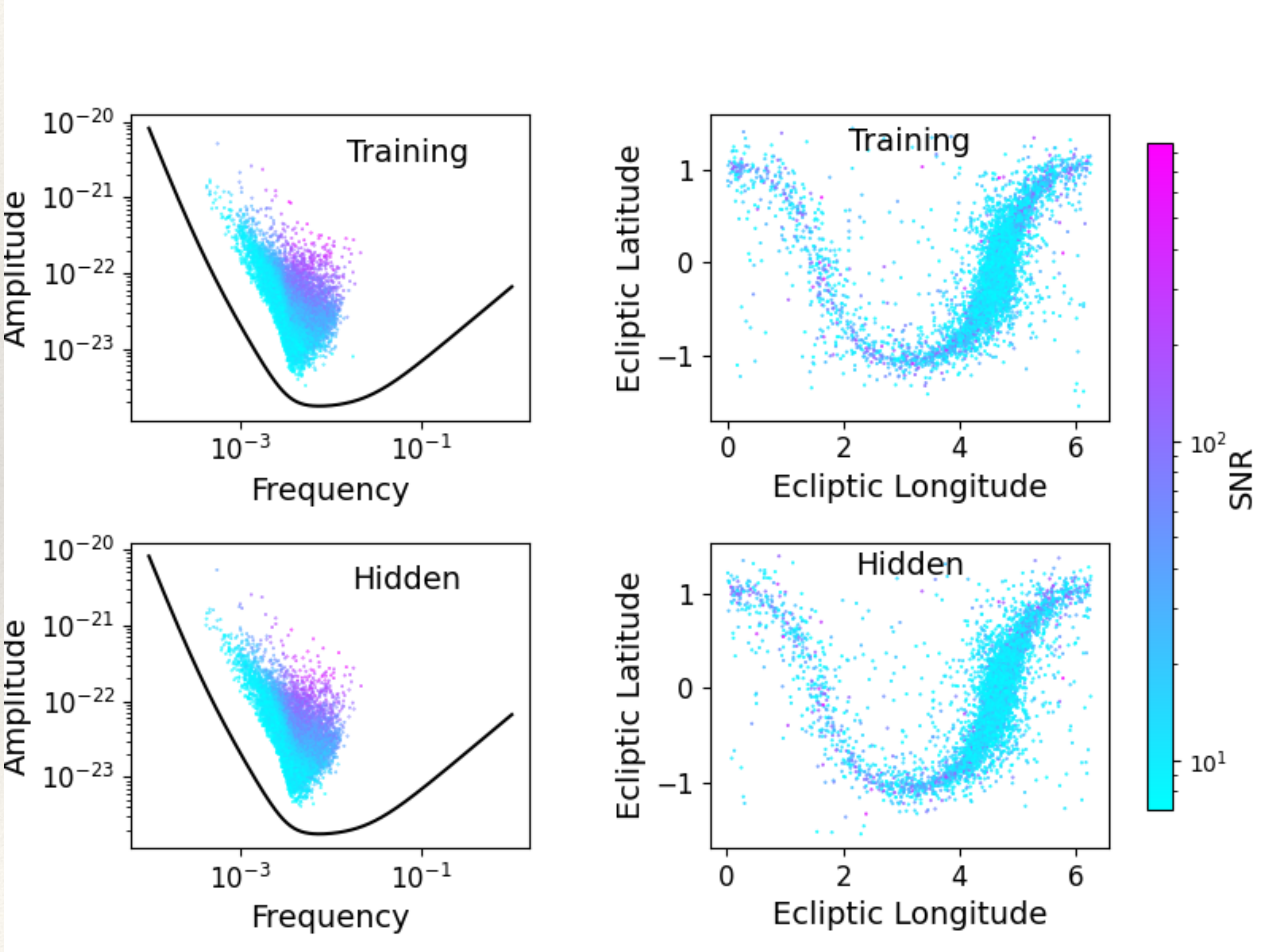
Strategy: white dwarf binaries

- White-dwarf binary signals are *compact in frequency*. Analysis updates binaries in frequency sub-bands in parallel. Number uncertain so use reversible jump. Tune proposals to improve efficiency (see Natalia's talk).



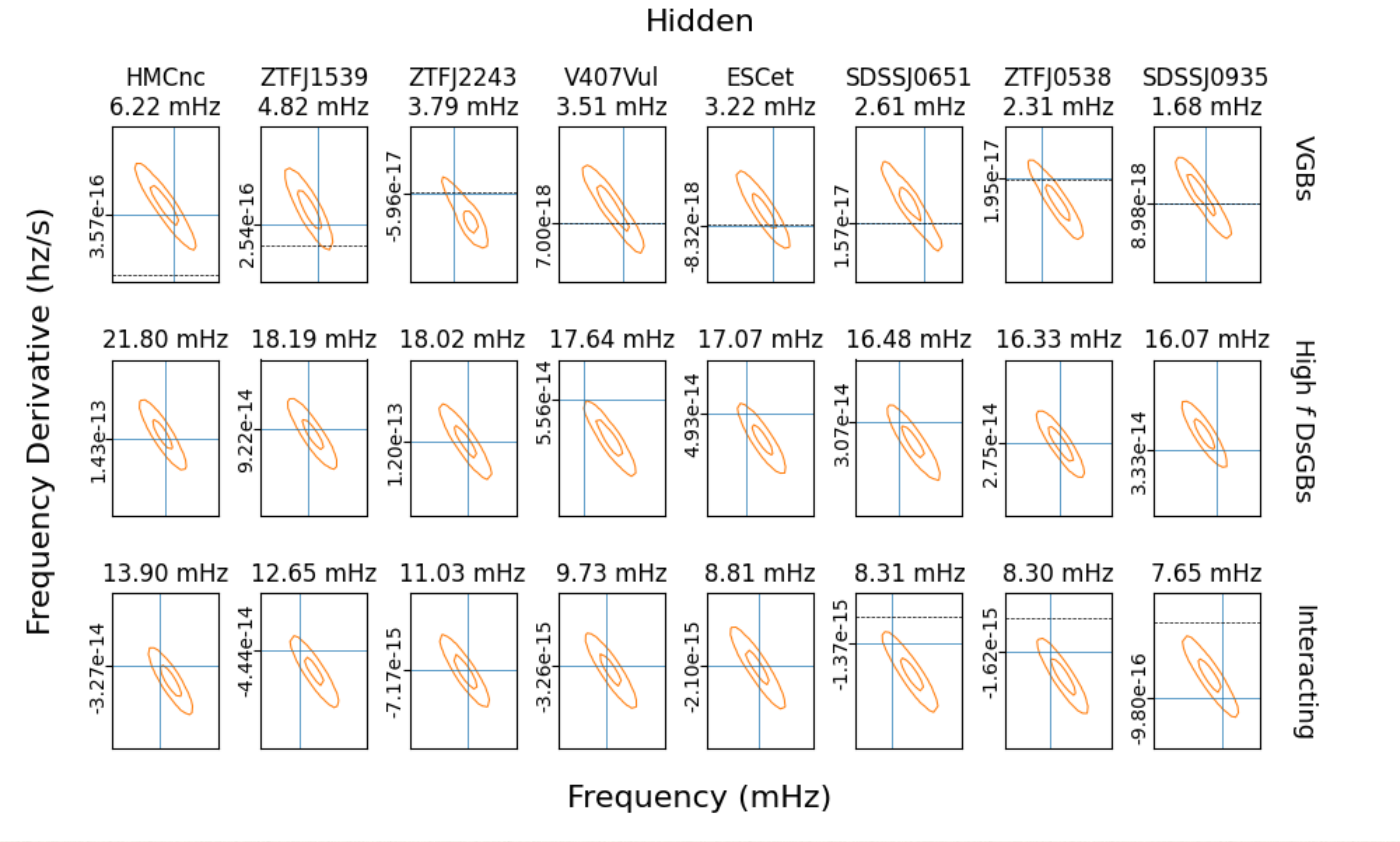
Results: white dwarf binaries

- Recover ~10,000 bright binaries distributed throughout the galaxy.



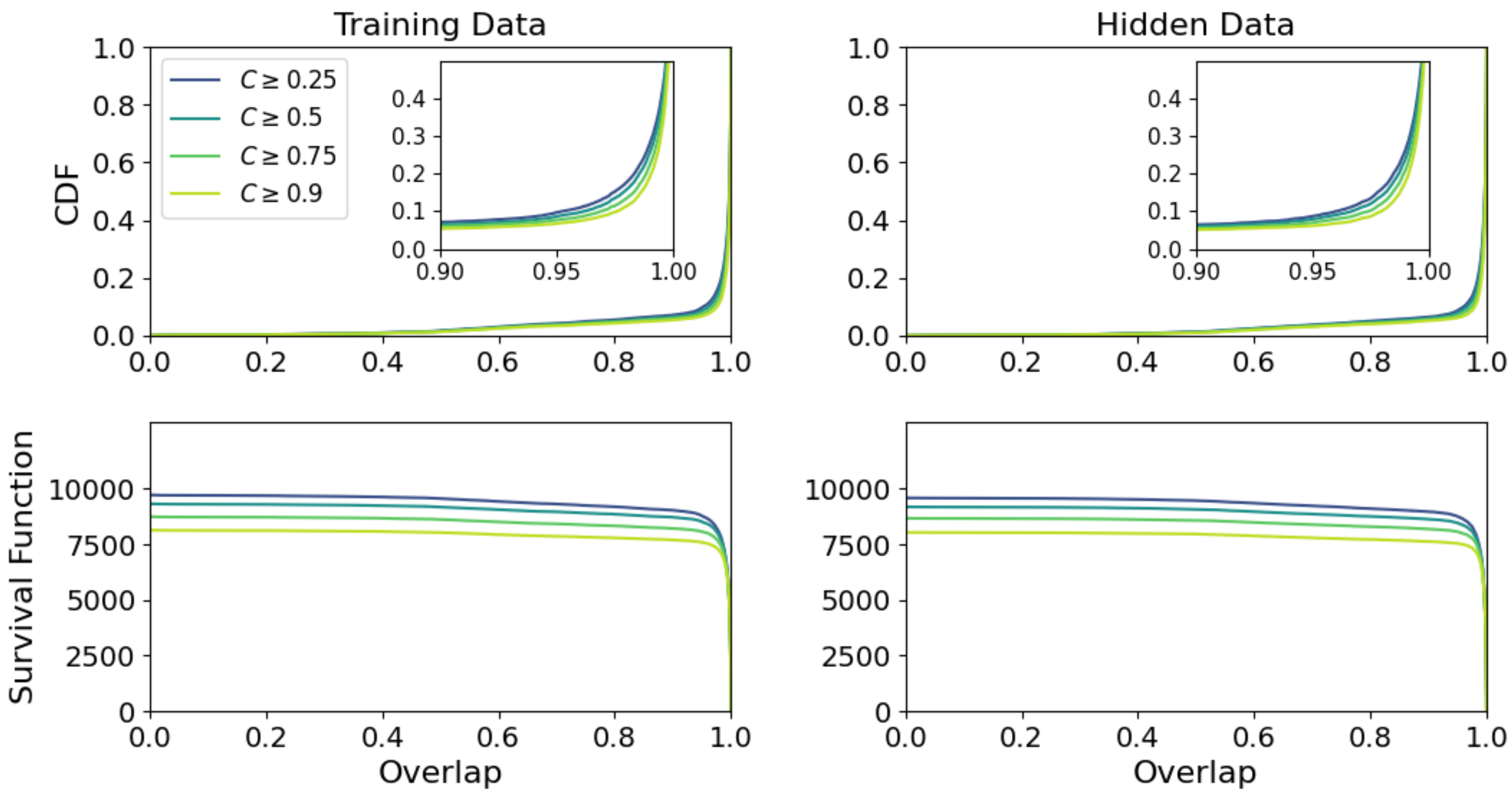
Results: white dwarf binaries

- Assess performance by looking at posteriors for white dwarf binaries known optically.



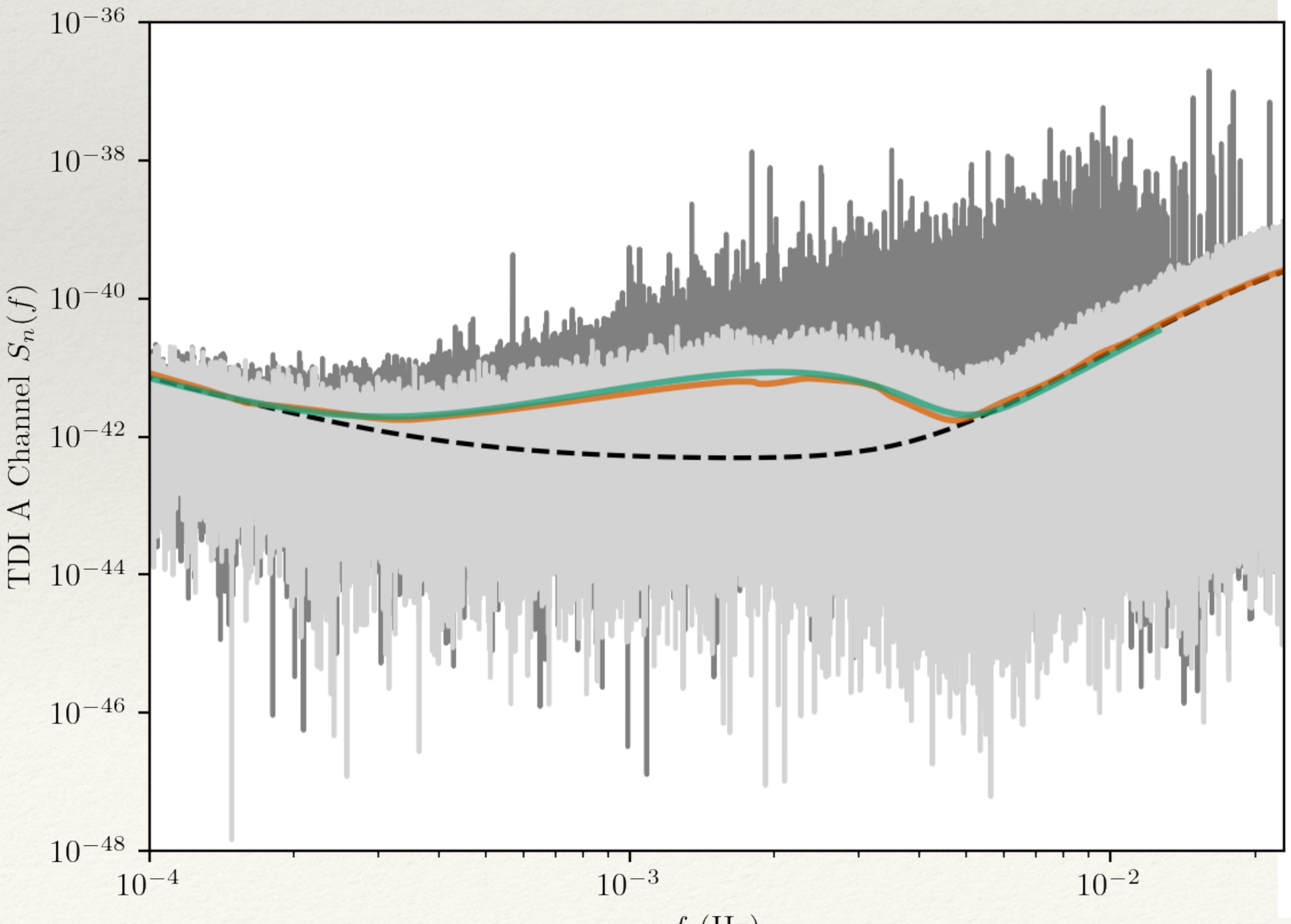
Results: white dwarf binaries

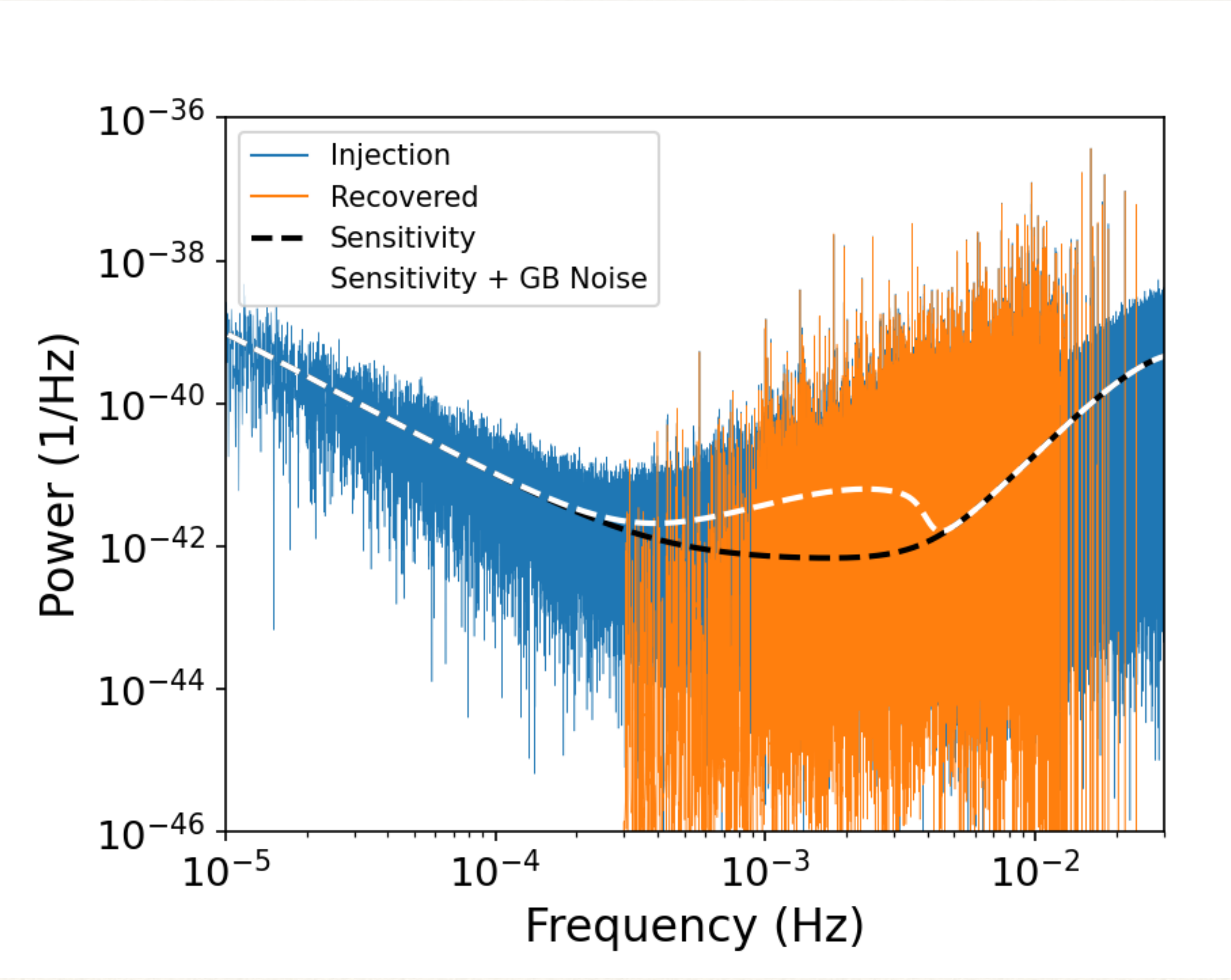
- ….and by comparing to the known injected catalogue.



Overall performance

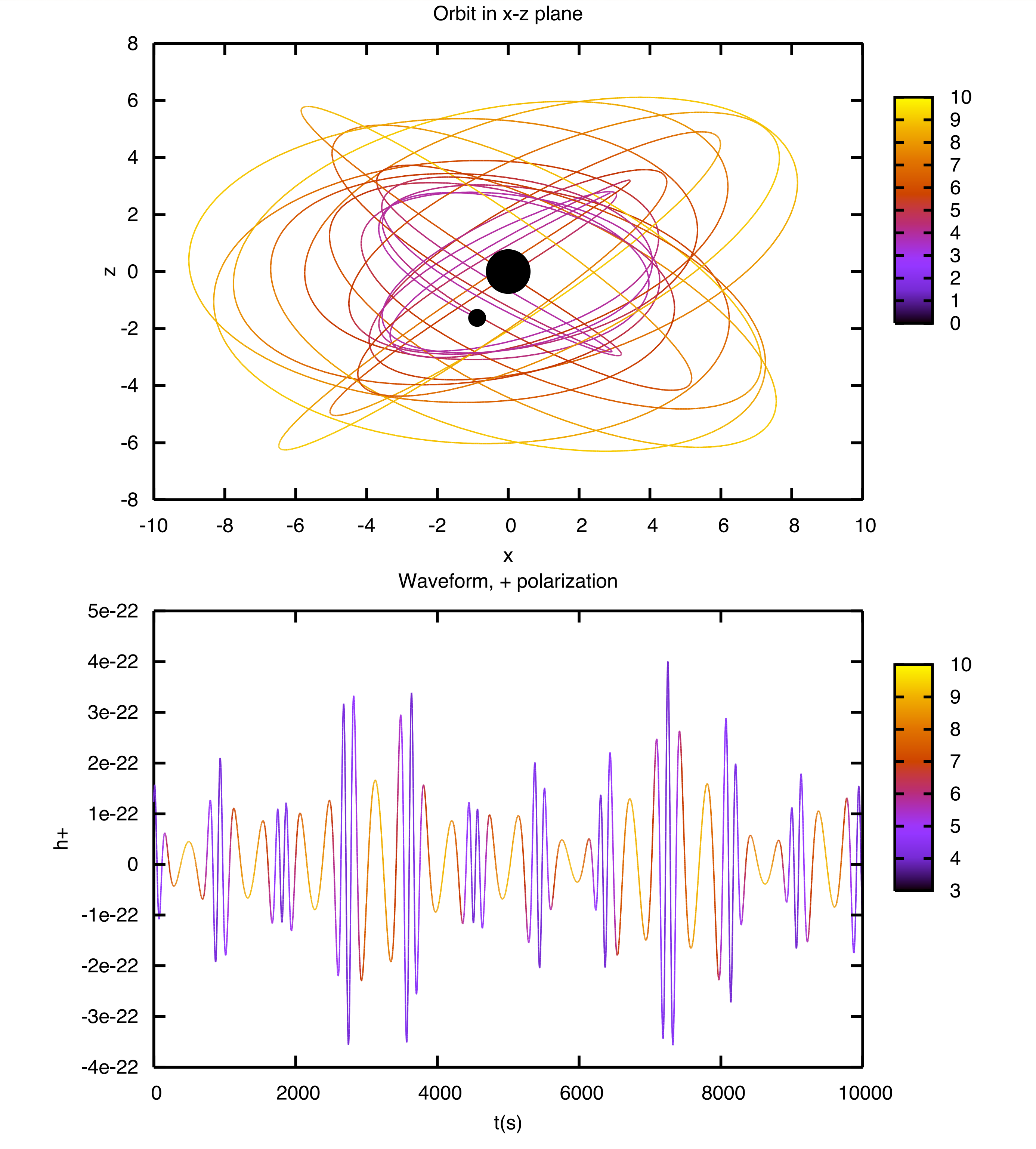
- In addition to MBH mergers and WD binaries, we fit the unknown noise level in the instrument, using a (stationary) parametric model.
- Four groups successfully analysed the *Sangria* data, with comparable levels of performance.
- Our approach required ~1 week on 4 GPUs.



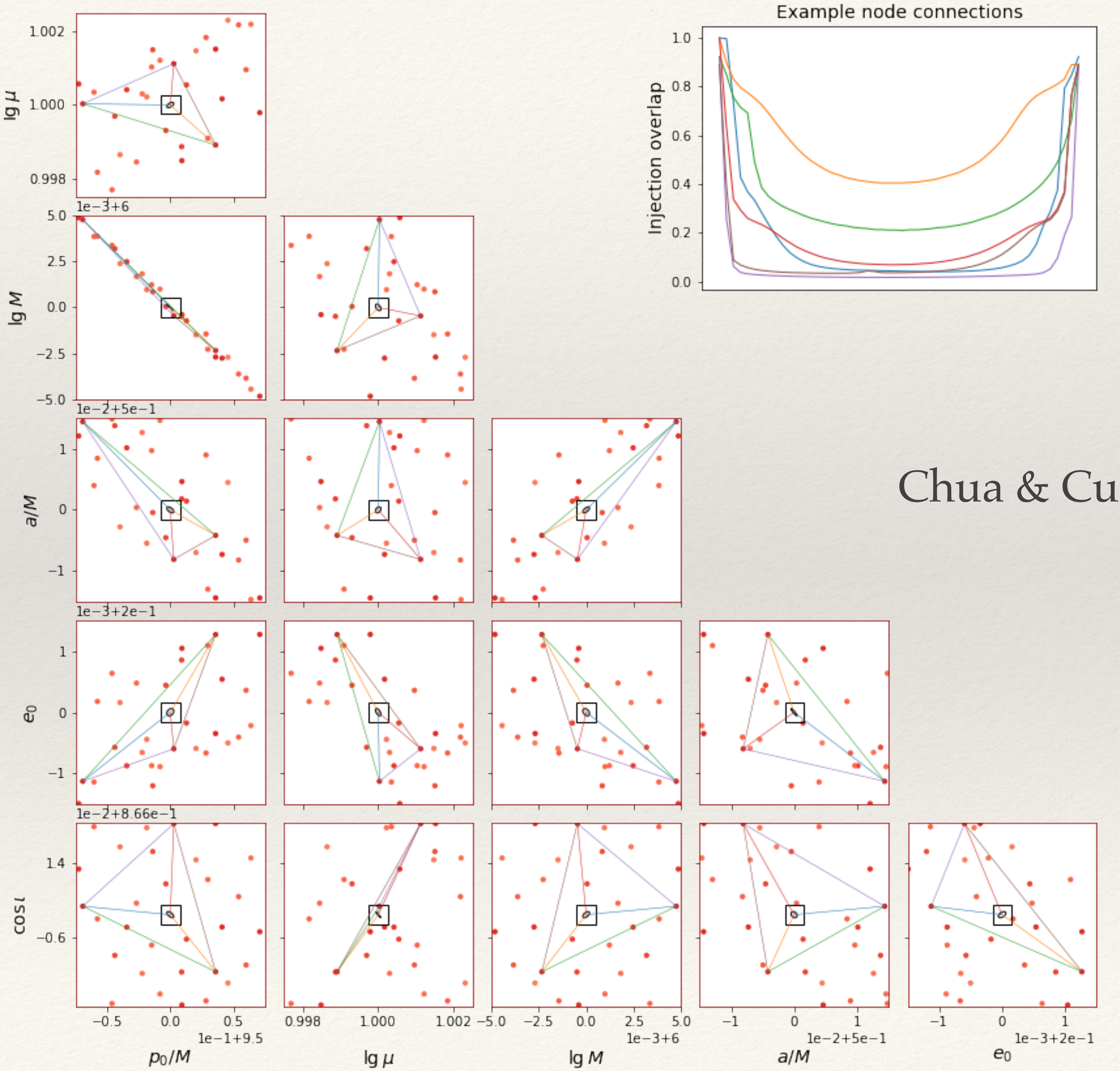


Outstanding challenges: EMRIs

- Various sources not yet included, including stellar-origin black hole mergers and EMRIs.
- EMRI waveforms show a rich structure built up from harmonics of three fundamental frequencies.
- EMRIs generate $O(10^5)$ cycles in strong field region close to central black hole.
- *In principle*: high precision measurements of system properties, including possible environmental effects and deviations from GR.
- *In practice*: narrow mode in big parameter space, many secondaries.



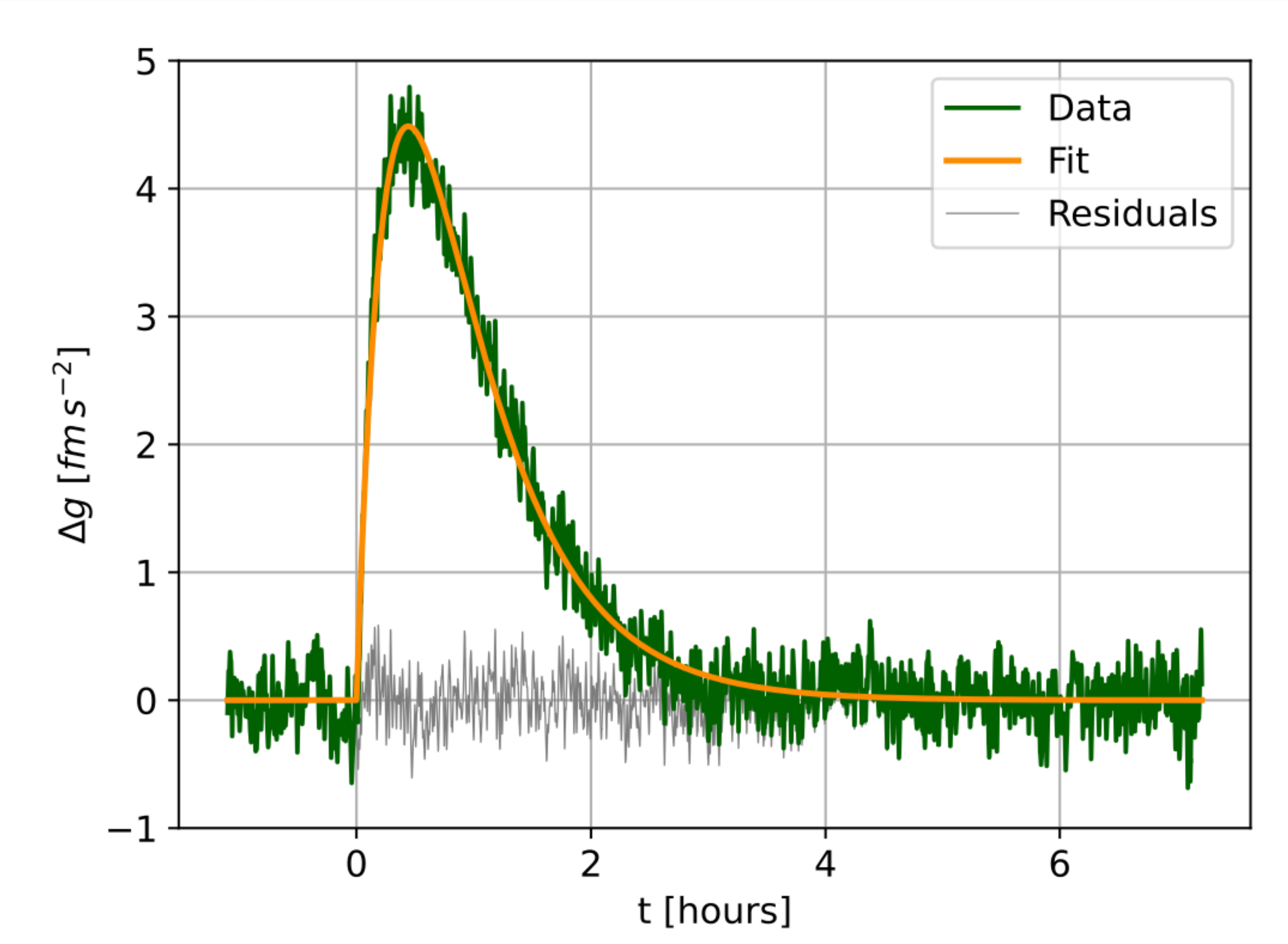
Outstanding challenges: EMRIs



Chua & Cutler (2021)

Outstanding challenges: glitches

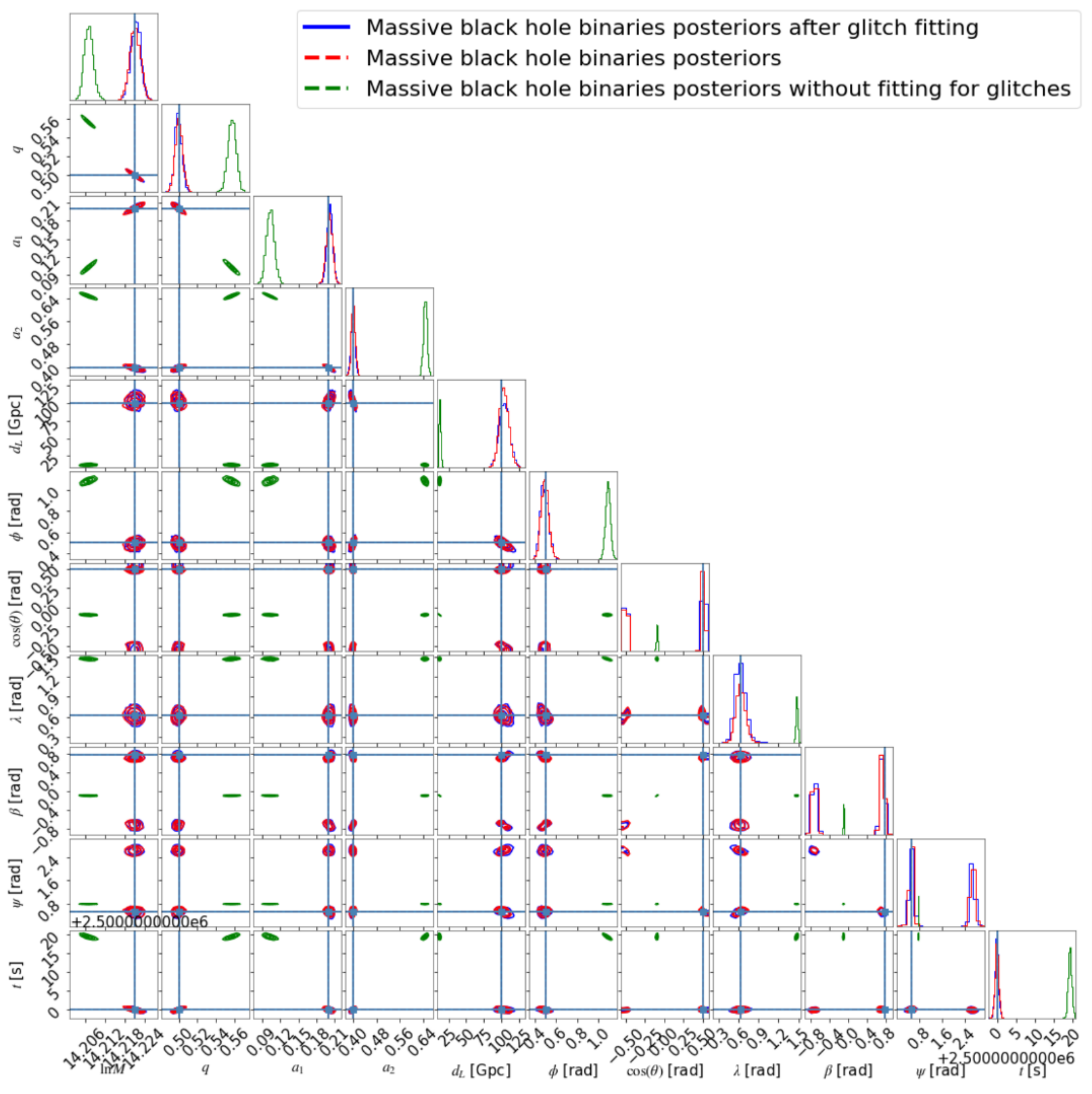
- LISA Pathfinder observed glitches at a rate of 1/day. Expect glitches in LISA too.
- Pathfinder glitches well described by a single exponential.
- No guarantee LISA glitches will have the same morphology.



$$h_1(t) = \frac{\Delta v}{\tau^2}\, t'\, e^{-t'/\tau}\, \Theta(t'), \quad t' = t - t_0$$

Outstanding challenges: glitches

- If a glitch overlaps an MBH merger, can get biases.
- Avoid biases by fitting for glitch simultaneously with signal parameters.
- Need reliable glitch model.
- But, glitches arise on spacecraft. So, at population level, glitches should follow a different distribution.
- Glitch fitting tested in the *Spritz* LDC data set.

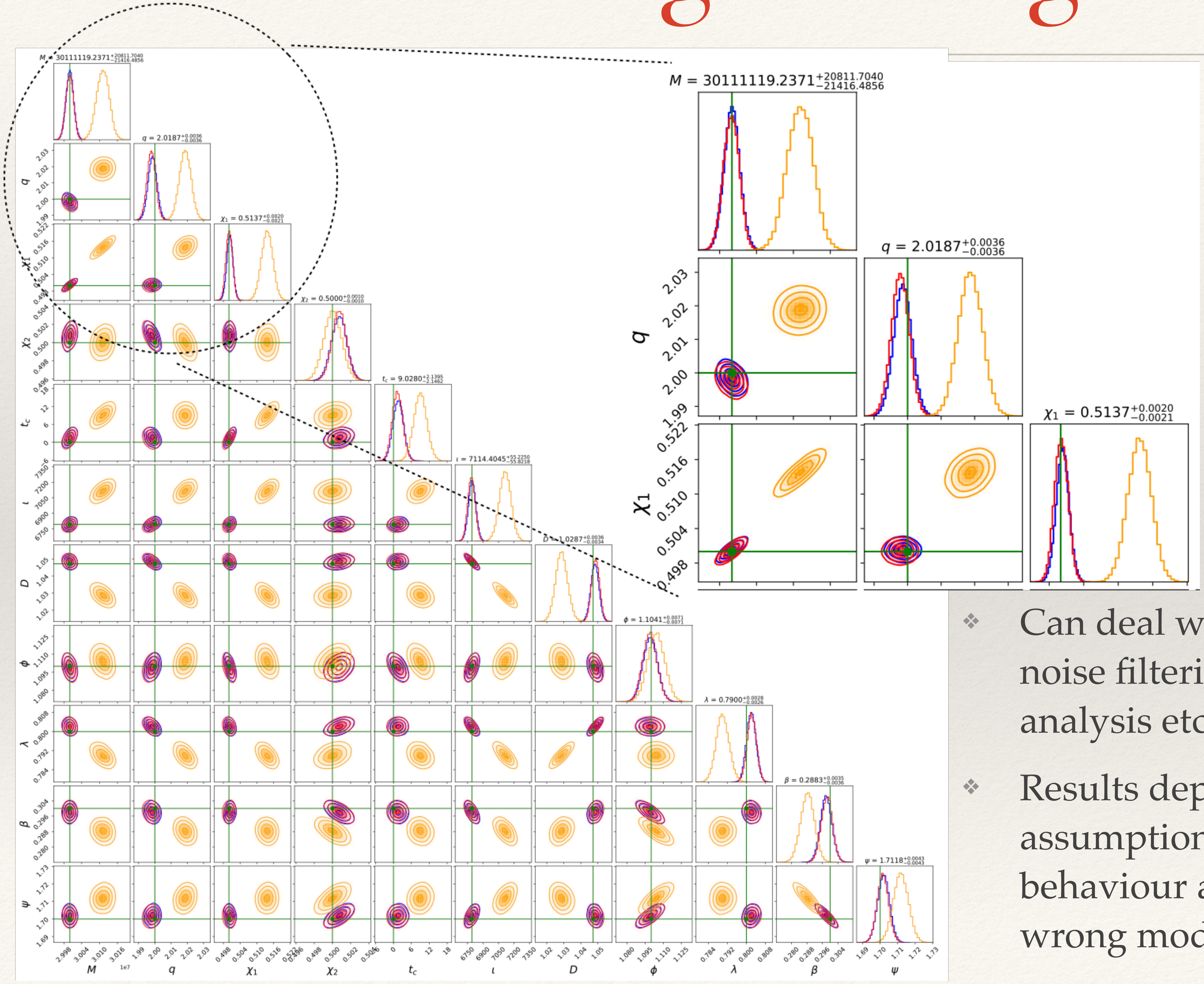


Outstanding challenges: gaps

- Many possible causes of gaps in the LISA data stream, of both known and unknown origin. Impact of antenna repointing gaps tested in *Spritz* data challenge.

Gap type	Frequency	Duration	Total loss (hr/yr)
Antenna repointing	every 2 weeks	3.3h	1%
PAAM angle adjust	3 per day	100s	0.3%
TM stray pot. est.	2/yr	1 day	0.56%
TTL coupling est.	4/yr	2 days	2.22%
Unplanned: platform	3/yr	2.5 days	2%
Unplanned: payload	4/yr	2.75 days	3%
Unplanned: micro-meteorites	30/yr	1 day	8%

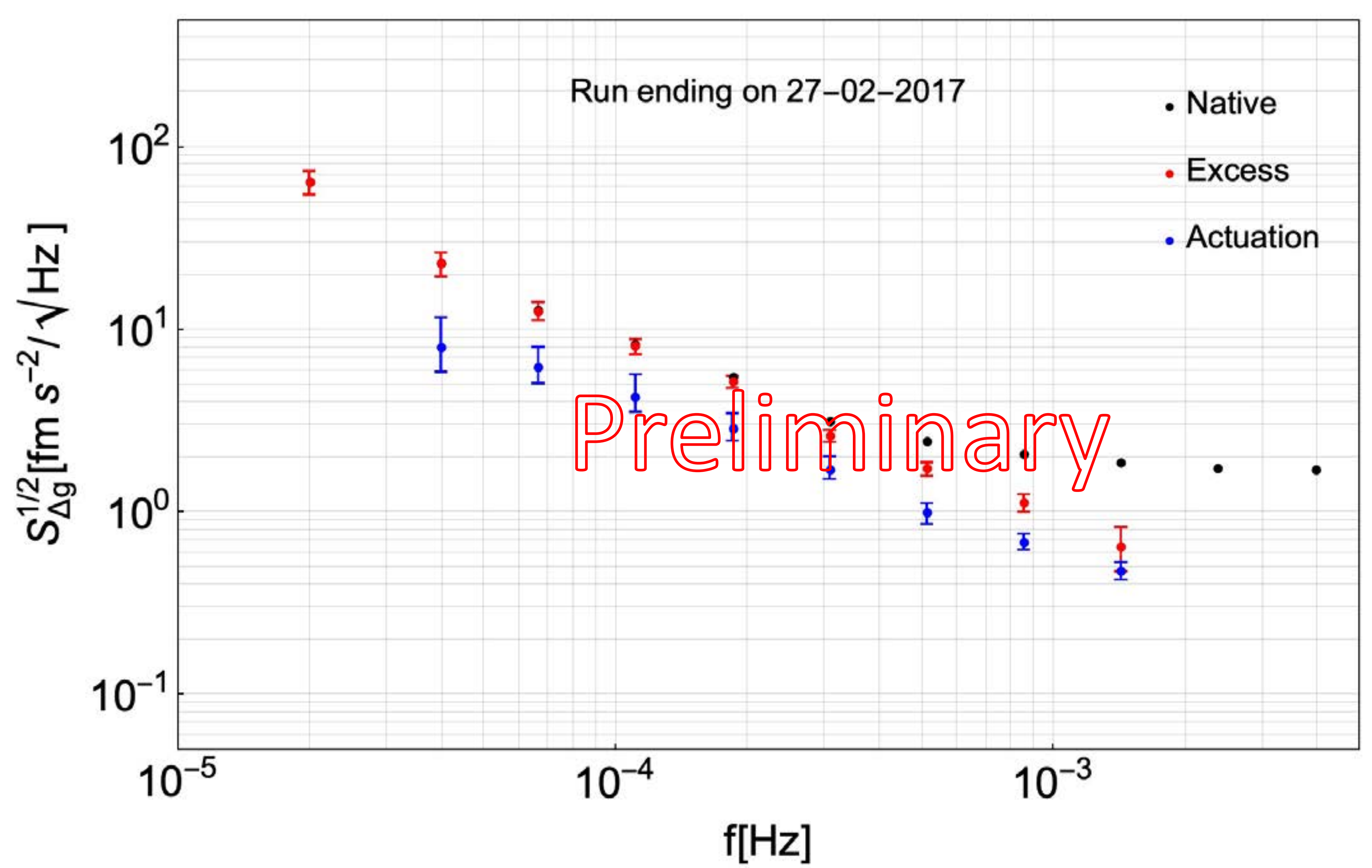
Outstanding challenges: gaps



- Can deal with gaps by gap filling, noise filtering, time-frequency analysis etc.
- Results depend critically on assumptions about noise behaviour across gap. Using the wrong model leads to biases.

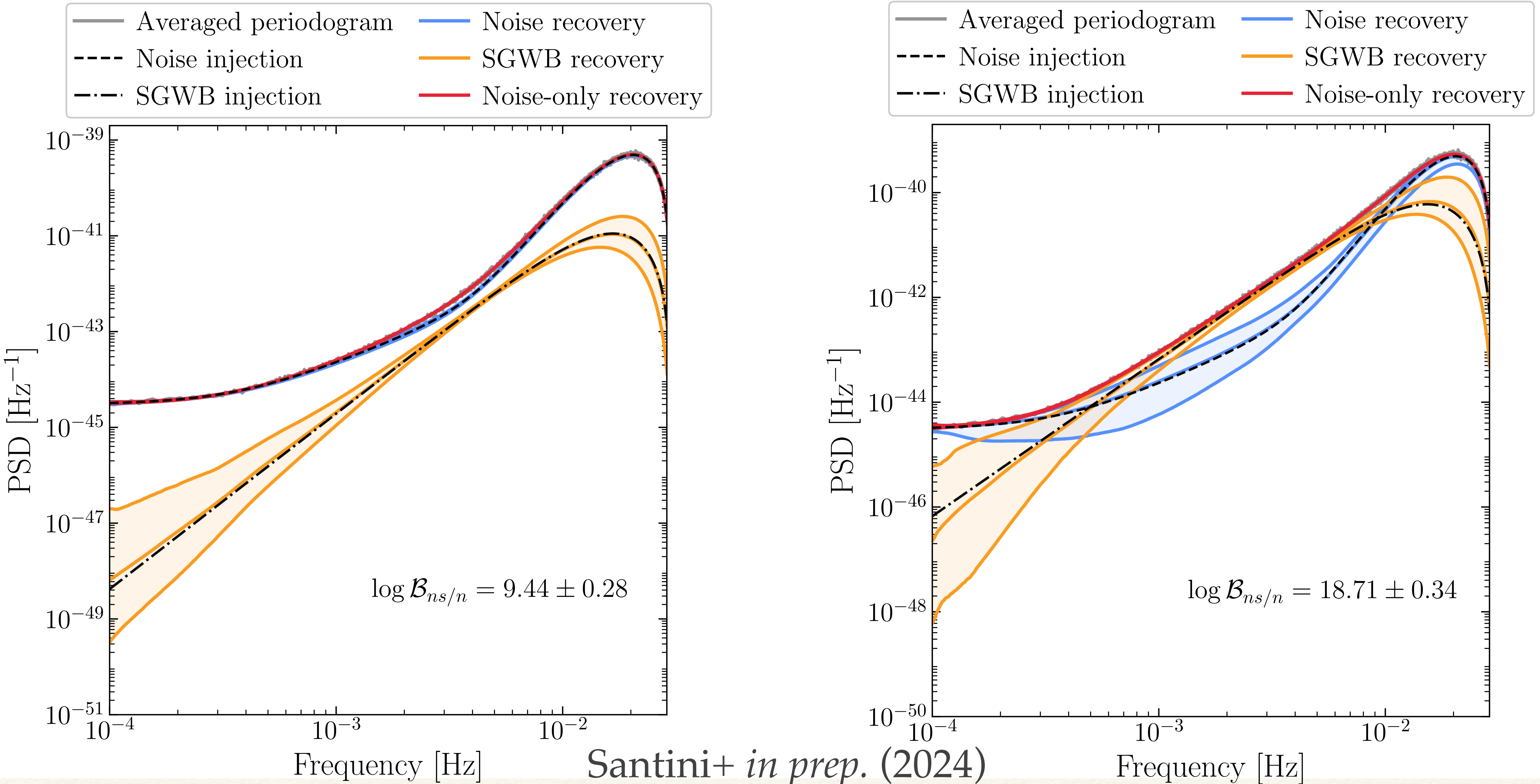
Outstanding challenges: lack of noise knowledge

- Bayesian approaches fit noise model. However, in LISA Pathfinder only 25% of total noise power was explained by measured noise sources.



Outstanding challenges: lack of noise knowledge

- At leading order, noise estimation and signal estimation are orthogonal, so PE for individual sources only modified by change in SNR, but problematic for backgrounds.
- Need flexible models to fit noise uncertainties (see Riccardo's talk).



Santini+ *in prep.* (2024)

Novel approaches: machine learning

- Machine learning is being used in various ways in data analysis for GW detectors
  - *searches*: use classifiers to identify potential signals in data from ground-based detectors;

Novel approaches: machine learning

- Machine learning is being used in various ways in data analysis for GW detectors
  - ***searches***: use classifiers to identify potential signals in data from ground-based detectors;
  - ***detector characterisation***: train networks to learn glitch morphologies and flag bad data periods;

Novel approaches: machine learning

- Machine learning is being used in various ways in data analysis for GW detectors
  - *searches*: use classifiers to identify potential signals in data from ground-based detectors;
  - *detector characterisation*: train networks to learn glitch morphologies and flag bad data periods;
  - *parameter estimation*: use a neural network to generate samples from the posterior distribution on the parameters of a source;

Novel approaches: machine learning

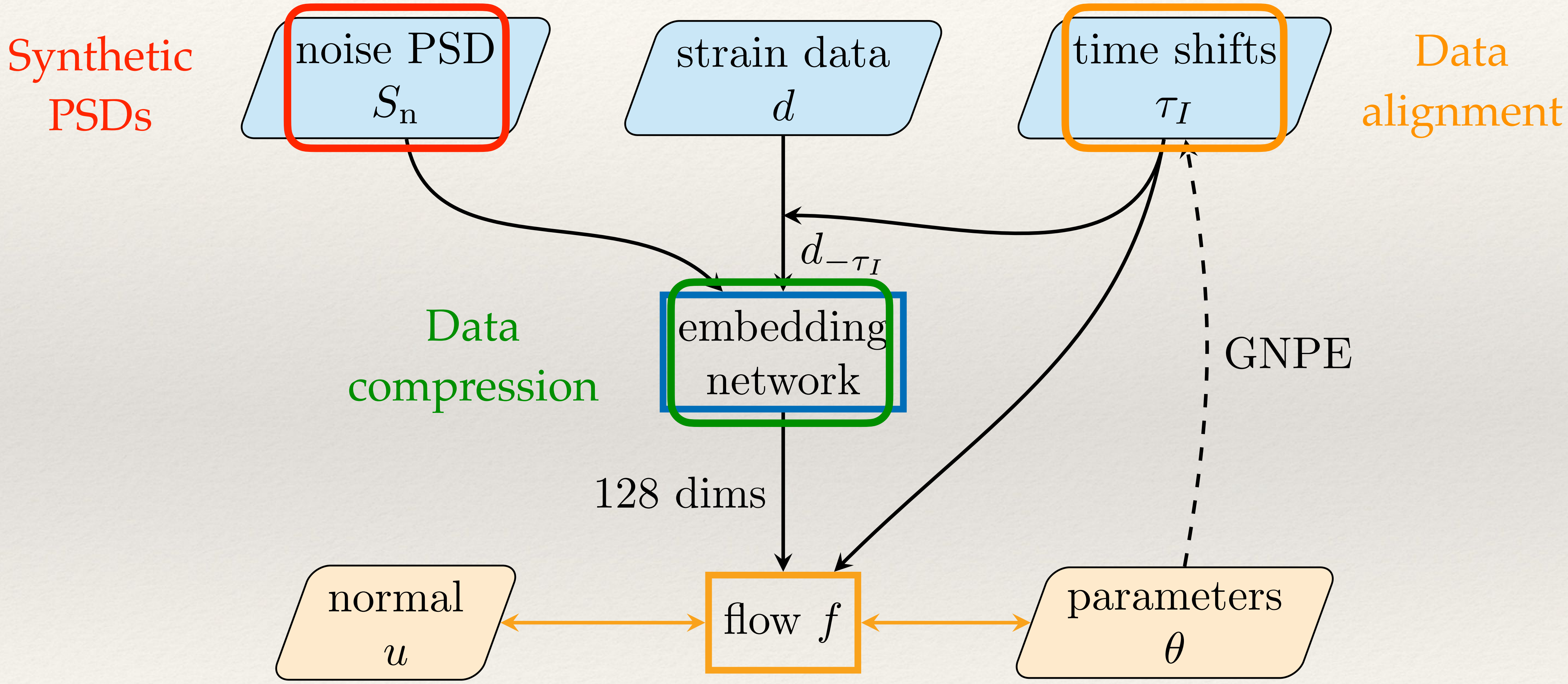
- Machine learning is being used in various ways in data analysis for GW detectors
  - ***searches***: use classifiers to identify potential signals in data from ground-based detectors;
  - ***detector characterisation***: train networks to learn glitch morphologies and flag bad data periods;
  - ***parameter estimation***: use a neural network to generate samples from the posterior distribution on the parameters of a source;
  - ***representation of distributions***: use neural networks to describe proposal distributions to use in sampling (see Natalia's talk).

Novel approaches: machine learning

- Machine learning is being used in various ways in data analysis for GW detectors
  - *searches*: use classifiers to identify potential signals in data from ground-based detectors;
  - *detector characterisation*: train networks to learn glitch morphologies and flag bad data periods;
  - *parameter estimation*: use a neural network to generate samples from the posterior distribution on the parameters of a source;
  - *representation of distributions*: use neural networks to describe proposal distributions to use in sampling (see Natalia's talk).
- Key challenges for LISA
  - *compression of input data*: need to project input data onto a suitable reduced representation to facilitate network training;
  - *overlapping sources*: data contains an unknown number of overlapping signals;
  - *high precision*: precise measurements means large training data sets.

Example: PE for LIGO using DINGO

- Train a *conditional normalising flow* that, when conditioned on observed data, generates samples from a density, $q(\theta|d)$, that approximates the true posterior, $p(\theta|d)$. Achieved by minimising *cross-entropy* on training set of simulated data.
- Various refinements needed to make it work in practice.



Synthetic PSDs

noise PSD S_n

strain data d

time shifts τ_I

Data alignment

$d_{-\tau_I}$

Data compression

embedding network

GNPE

128 dims

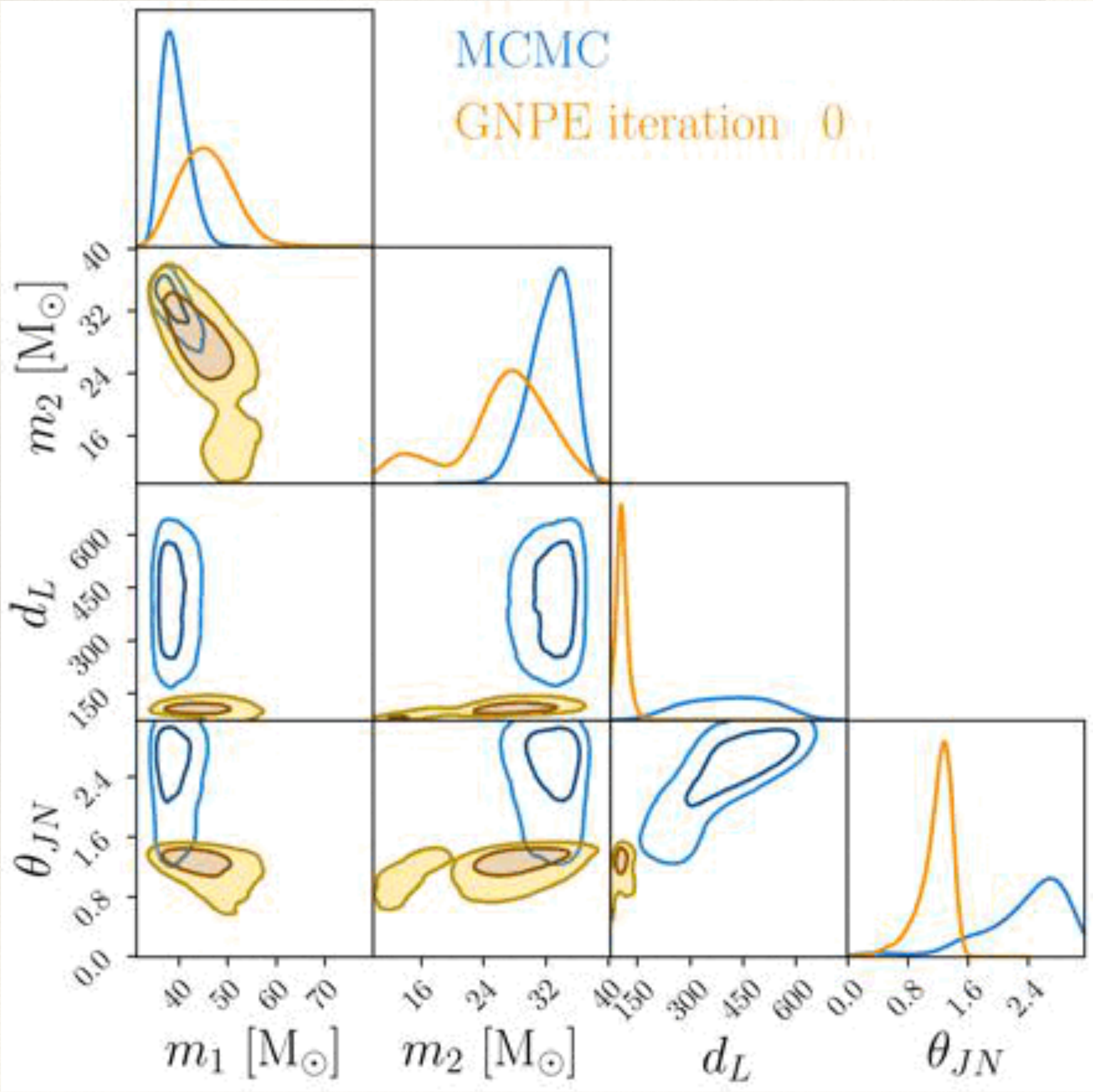
normal u

flow f

parameters θ

Example: PE for LIGO using DINGO

- DINGO posteriors for GW events indistinguishable from standard sampling, but much faster.
- Related techniques have been applied to LISA measurements of stochastic GW backgrounds.
- Extension to LISA MBH mergers currently in development.
- Possible LISA applications:
  - low latency alerts;
  - provide initial parameter estimates to global fit;
  - catalogue representation?
  - replace whole global fit?
- Simulation-based-inference would be a natural approach to tackle instrumental complexities.



Summary

- Currently operating facilities are observing gravitational waves in the 1-1000Hz (LIGO/Virgo) and nanohertz (PTA) bands.
- LISA will open up the millihertz band, which is expected to be very rich in sources, including: *ultracompact binaries* in the Milky Way, *massive black hole mergers*, *extreme-mass-ratio inspirals*, *stellar-origin black hole mergers* and *stochastic backgrounds* generated in the early Universe.
- LISA data analysis is a big model problem, requiring simultaneous fitting of a large number of overlapping sources of many different types.
- Progress is being made using classic *stochastic sampling* methods, augmented with *reversible jump*, *affine-invariant sampling* and *parallel tempering*.
- Several problems still need to be overcome, including simultaneous treatment of *instrumental artefacts* (gaps, glitches and uncertain noise) and the *search and characterisation of EMRIs* and *SOBHs*.
- Machine learning approaches to LISA data analysis are being explored and have potential applications to low latency, search and to accelerate the convergence of existing algorithms.