University of Glasgow

GR-Amaldi – July 14th, 2025

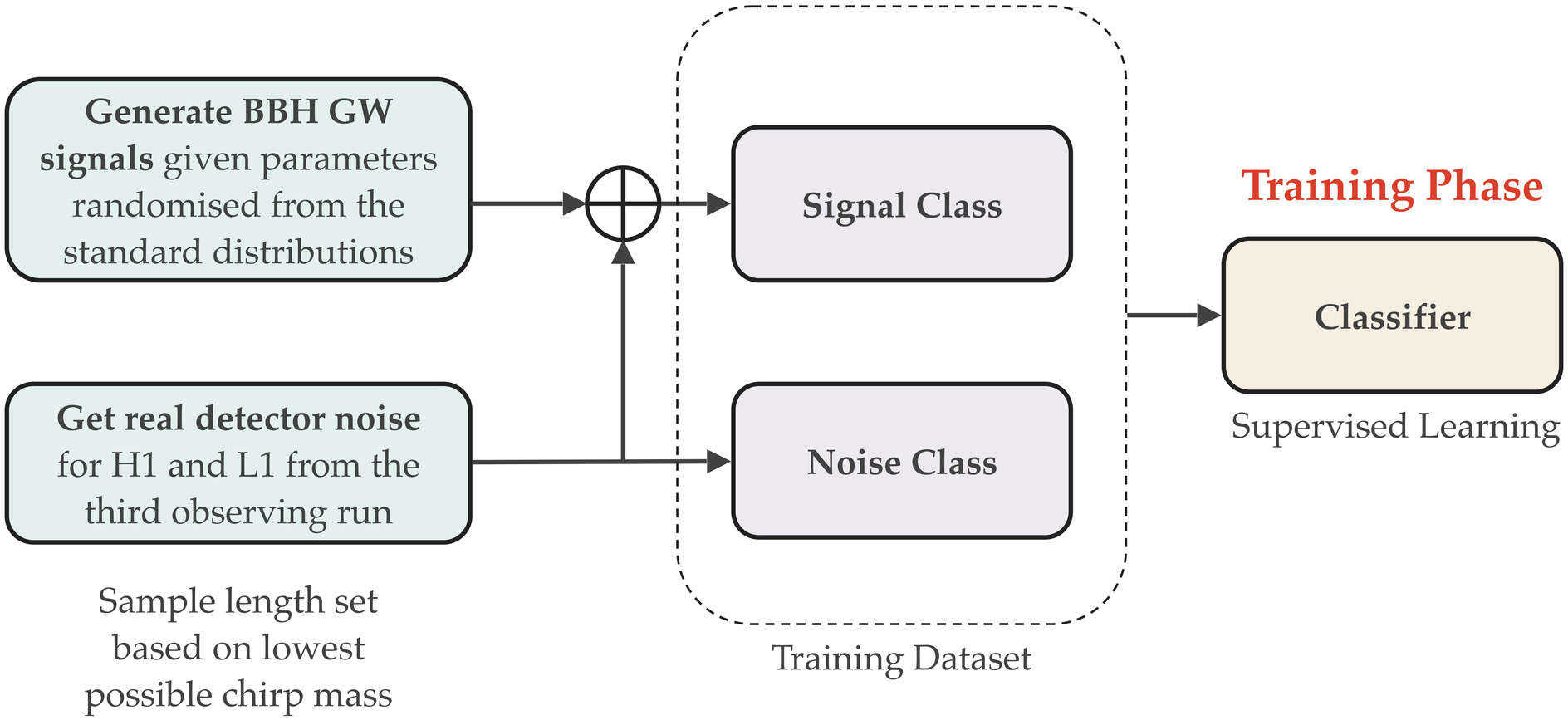
Identifying and Mitigating Machine Learning Biases for the Gravitational Wave Detection Problem

Narenraju Nagarajan
Chris Messenger

arxiv: 2501.13846 (accepted for publication in PRD)

IGR

Problem Setup: Detection of BBH GWs interpreted as a classification problem in the deep learning framework.



Motivation: Previous works were affected by ML biases. Finding the source of these biases and mitigating them might allow ML to move closer to the detection performance of matched-filtering pipelines.

M. B. Schäfer et al., Phys. Rev. D 107, 023021 (2023) → Reported bias against **long duration signals** and inability to reject **non-Gaussian noise artefacts**

Koloniari, Alexandra E. et al. (2024) -- AresGW → Reported bias in the **tails of the chirp mass distribution**

Marx, Ethan et al. (2024) -- AFrame → Reported bias against **lower chirp mass** and **high mass ratio signals**

What is Bias in this context?:

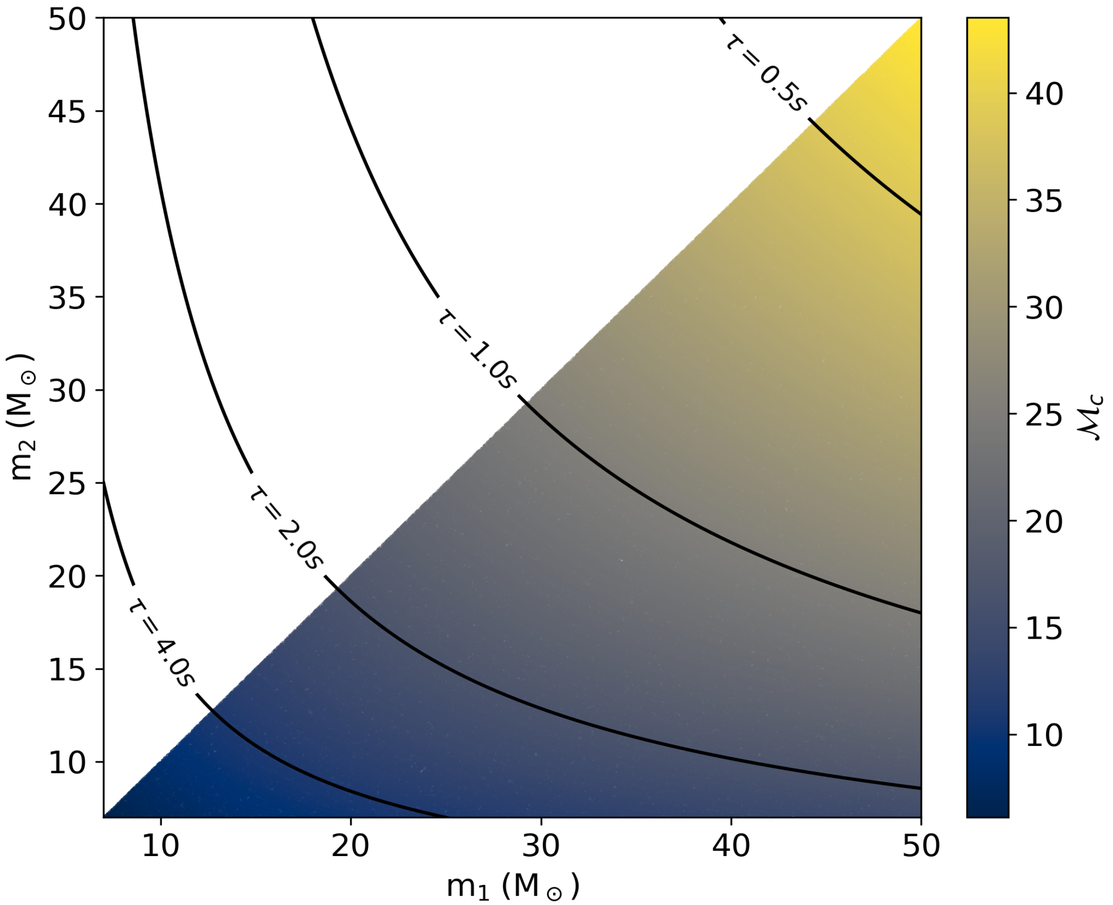
Identifying and Mitigating Machine Learning Biases for the Gravitational Wave Detection Problem

- Imagine your definition of a “perfect” search/classification analysis
  - This is likely to be either a Bayesian of Frequentist or other?
  - Remember, an ideal & realistic analysis **does not detect everything**
- We don’t have access to an ideal analysis for comparison
  - So biases in detection e.g. as a function of chirp mass are measured relatively

Bias due to limited samples

Imbalance in the sample distribution for a certain GW parameter might introduce biases.

Vokinger, K.N et al. Commun Med 1, 25 (2021)
Tang, Kaihua et al. ArXiv abs/2009.12991 (2020)



$$\tau = t_c - 5(8\pi f_l)^{-8/3}\mathscr{M}_c^{-5/3}$$

Create image Clock with time set to 5:40

Image created

The time is 10:10 on that image. I asked for 5:40. Correct this.

Image created

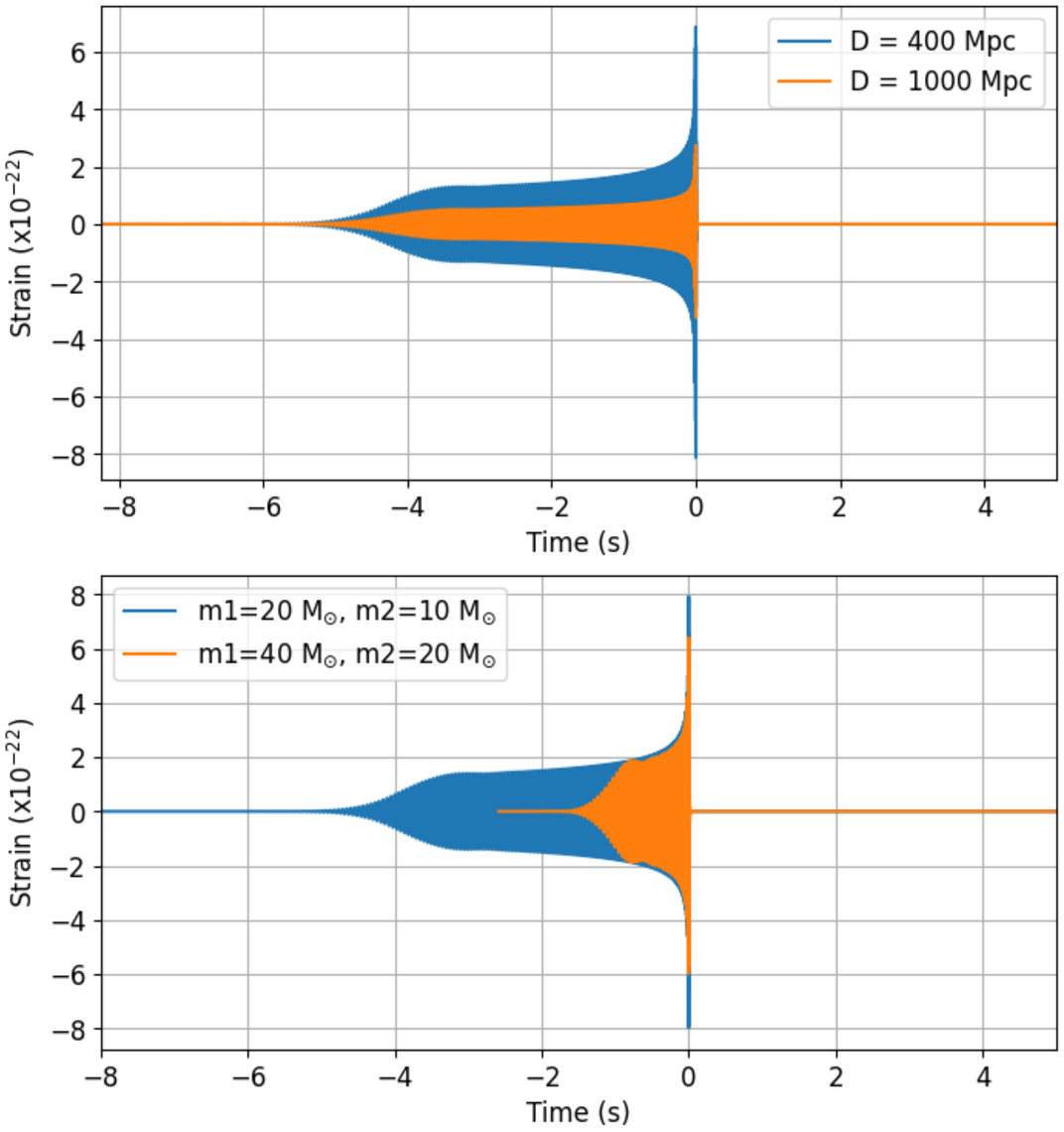
Is it true that chatGPT cannot create images of a clock with times other than 10:10?

No, that's not true — ChatGPT *can* create images of clocks showing any time, including times other than 10:10. However, some AI image generation systems (including DALL·E, which powers image generation here) sometimes default to 10:10 due to training data biases, as that's a common time shown in product photos of clocks. It's not a hard limit, and I can correct it on request — as I did when generating the 5:40 clock in the last image.

Bias due to lack of variation

Lack of augmentation along GW parameters that the model cannot interpolate trivially or lack of variation in noise characteristics.

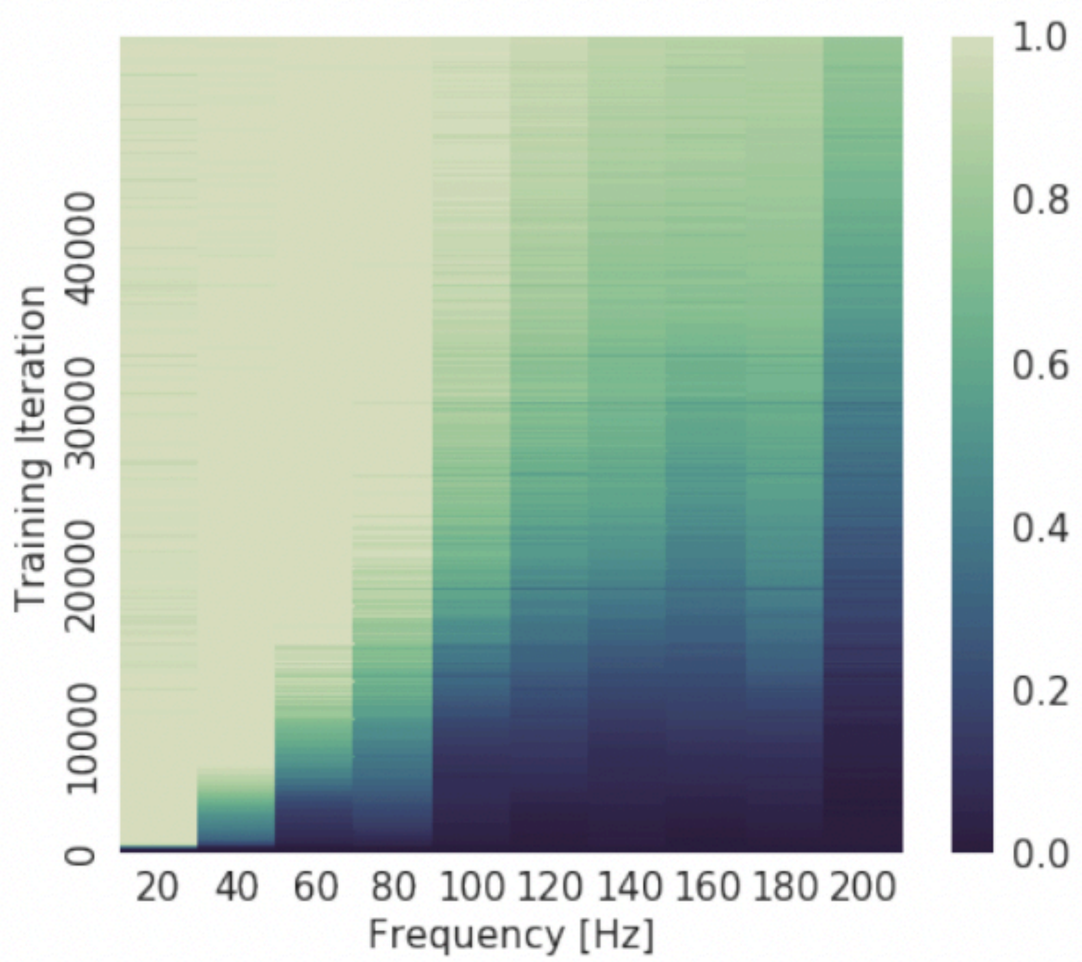
Shah, Harshay et al. ArXiv abs/2006.07710 (2020)
A. Mikołajczyk et al., IIPhDW, pp. 117-122 (2018)



Spectral bias

Lower frequency features are easier to learn than higher frequency features. Generalisation toward higher frequencies are typically brittle.

Rahaman, Nasim et al. ICML (2018).
Cao, Y. et al., arXiv:1912.01198 (2019)



ML model was trained to learn sine waves at different frequencies but same signal duration. The colours represent how well the network learns samples of a certain frequency during training. 1 is perfect and 0 is non-existent learning.

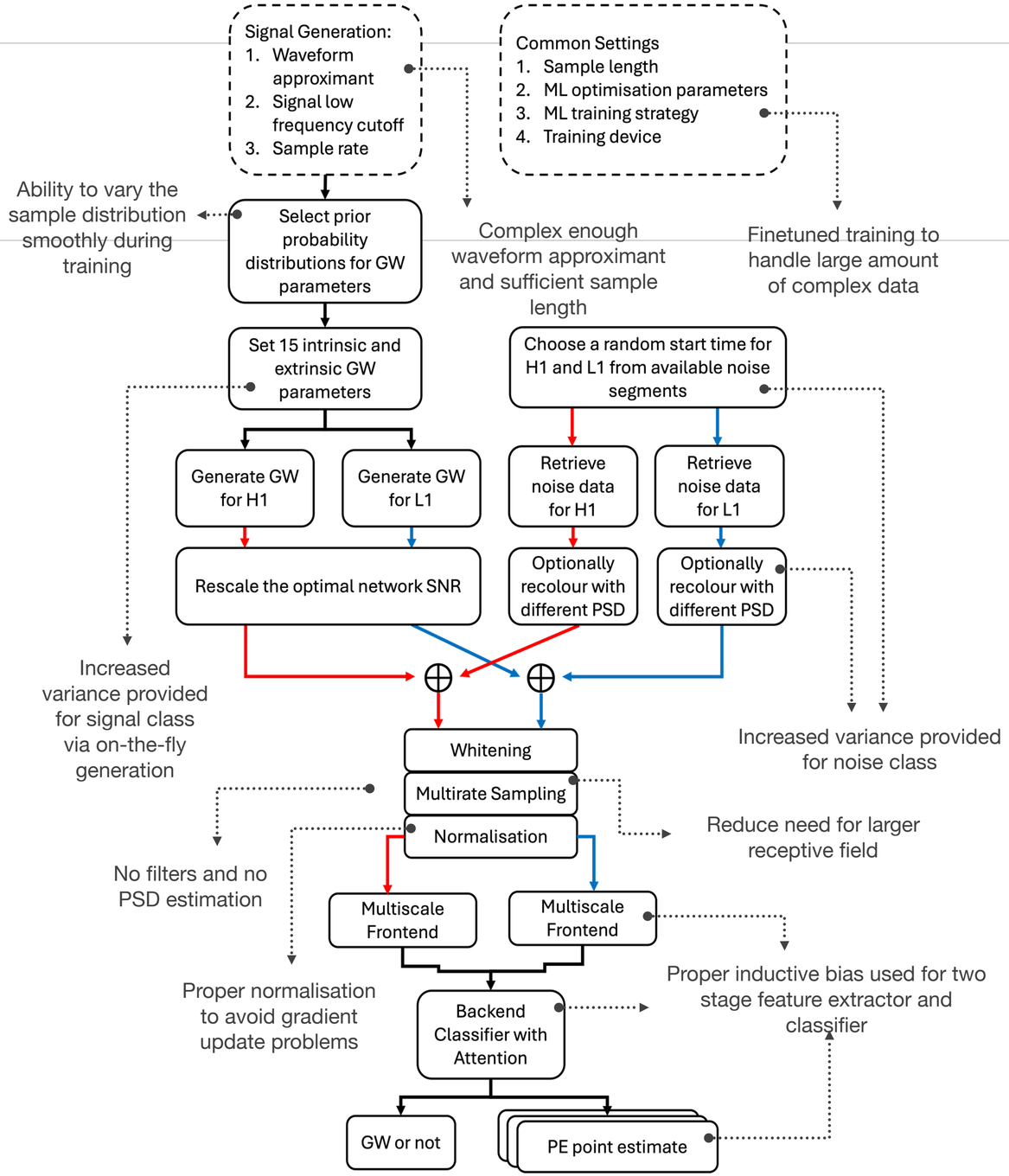
Rahaman et al. Proceedings of the 36th International Conference on Machine Learning, PMLR 97:5301-5310, 2019.

Other biases discussed in the paper [arxiv: 2501.13846 - accepted for publication in PRD]

1. Bias due to class overlap
2. Bias due to class imbalance
3. Bias due to train-test dataset overlap
4. Bias due to train-test mismatch
5. Bias due to disproportionate evaluation
6. Bias due to limited feature representation
7. Bias due to sample difficulty
8. Bias due to insufficient information

and possibly more…

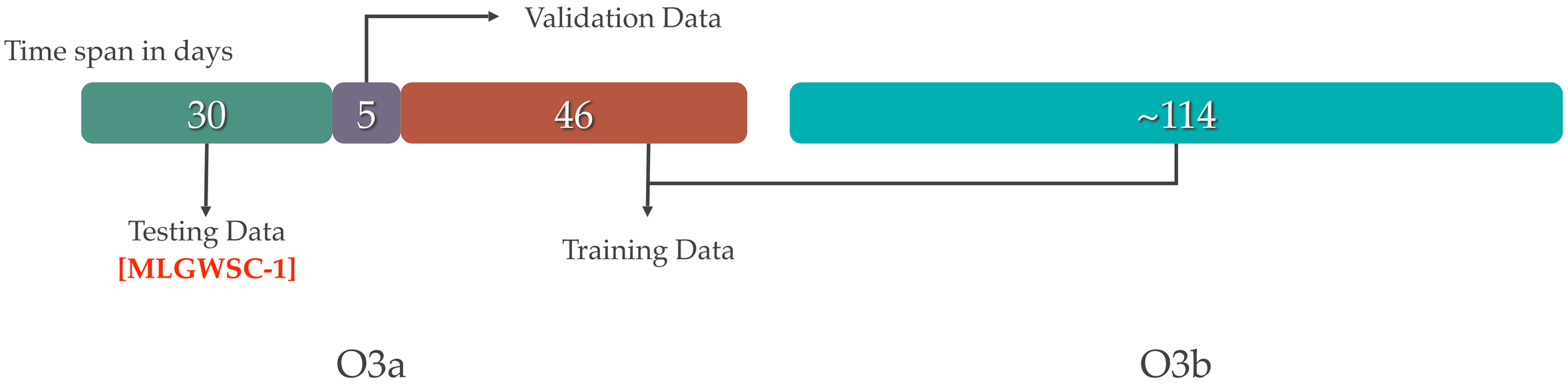
Sage Methodology

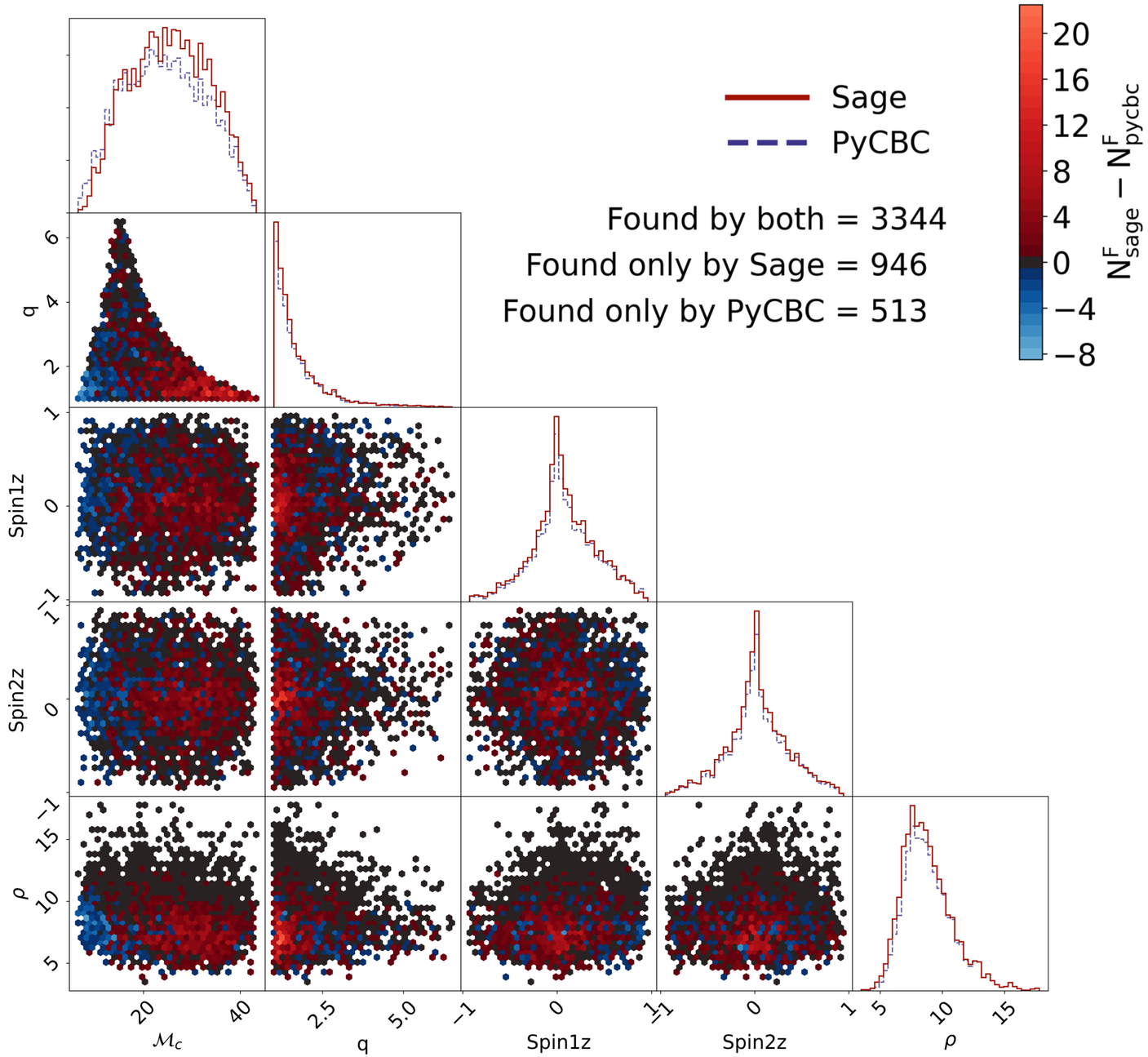


- The primary features of SAGE
  - Bias is very sensitive to the mass training distribution
  - Thoughtful network design addresses inductive bias
  - Out-of-distribution PSDs aid classification
  - Training on real data is key

Our aim was to use the MLGWSC-1 [Schäfer *et al.*, 2023] dataset for apples-to-apples comparison

1. Injection study with 1 month of O3a detector noise (H1, L1)
2. Injected signals generated using IMRPhenomXPHM
3. Mass priors are $U[7, 50]$ solar masses with $m1 > m2$
4. Spins χ are distributed isotropically with magnitudes between $[0.0,\ 0.99]$
5. Other GW parameters were randomised from the standard distributions

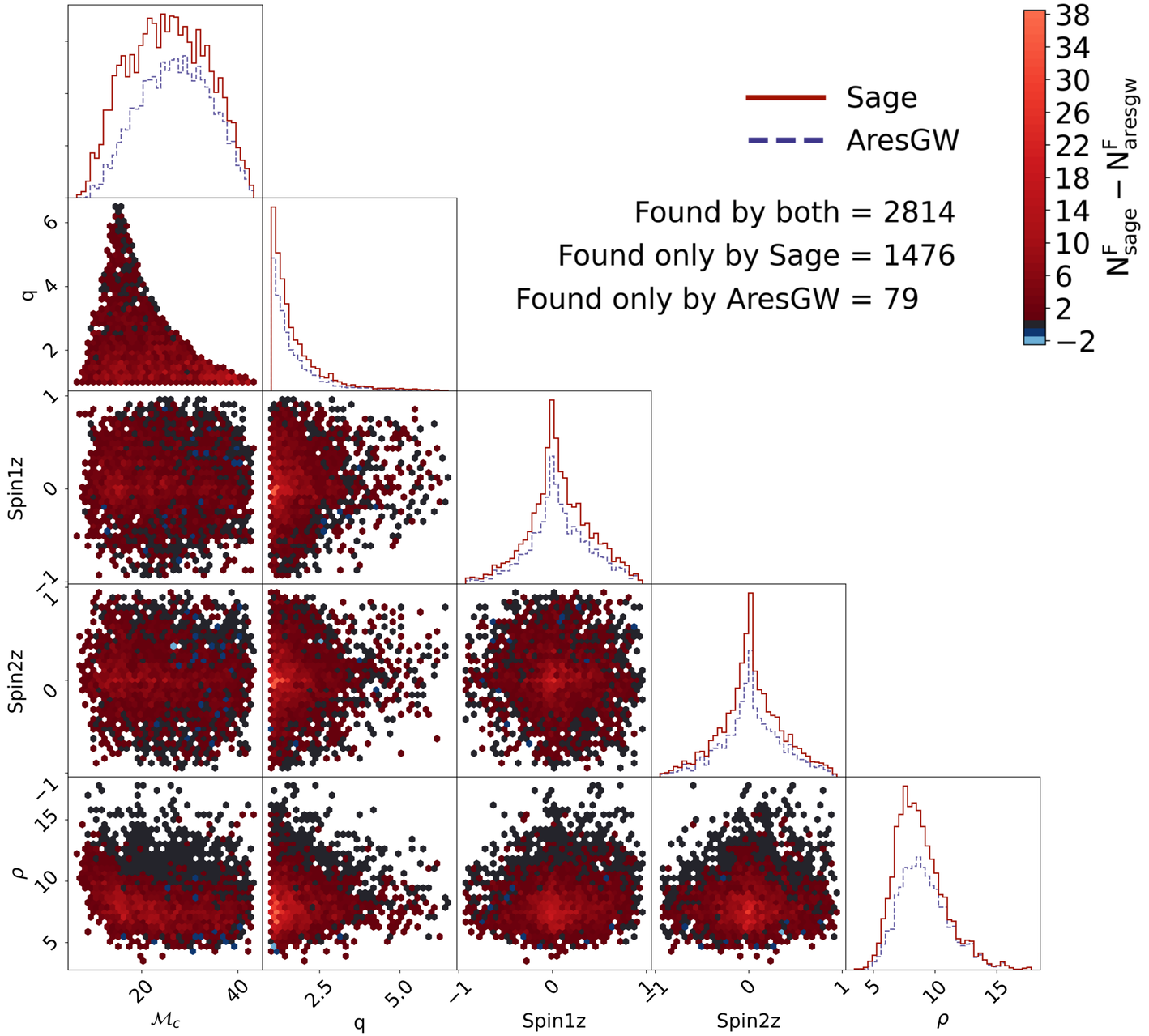




Comparing the histogram of found injections between Sage and PyCBC for the injection study at a False Alarm Rate (FAR) of 1 per month.

We find 443 signals more than matched-filtering and is equivalent to an increase of ~11.2% in the number of signals detected.

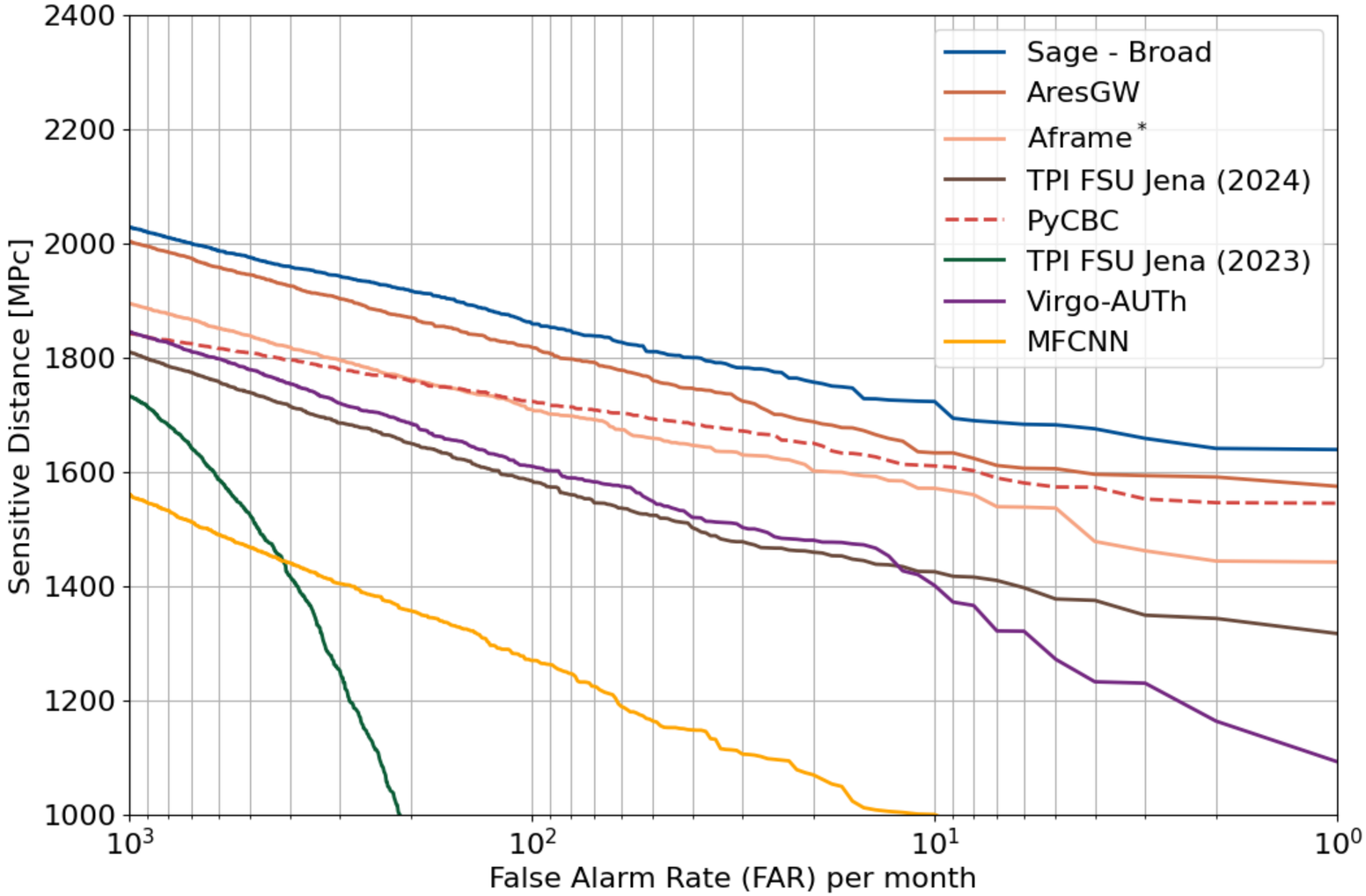
PyCBC results obtained from - M. B. Schäfer et al., Phys. Rev. D 107, 023021 (2023)



Comparing the histogram of found injections between Sage and AresGW for the injection study at a False Alarm Rate (FAR) of 1 per month.

We find 1397 signals more than AresGW on this dataset and is equivalent to an increase of ~48.29% in the number of signals detected.

AresGW results obtained from - P. Nousi et al., Phys. Rev. D 108, 024022 (2023)



$$V(\mathcal{F}) \approx \frac{V(d_{max})}{N_I} \sum_{i=1}^{N_{I,\mathcal{F}}} \left(\frac{\mathcal{M}_{c,i}}{\mathcal{M}_{c,max}} \right)^{5/2}$$

Parameter-dependent sensitive distance, $\propto V(\mathcal{F})^{\frac{1}{3}}$, as a function of false alarm rate per month.

SAGE (Empirical Contributions)

- PSD estimation is not required for ML-based detection

- Mitigating biases might aid in glitch rejection

- OOD PSDs are good for classification

- Data/parameter efficiency via GW domain knowledge

SAGE

arxiv: 2501.13846 - accepted for publication in PRD

Identified sources of biases in ML for GW detection problem

Provided training strategies and mitigation tactics to address biases concurrently.

We detect ~11.2% more signals than the benchmark PyCBC result via the MLGWSC-1 injection study at an FAR of 1/month in O3a noise.

Analysis of O3 is underway :)

University of Glasgow

Talk @ AIslands – May 13th, 2025

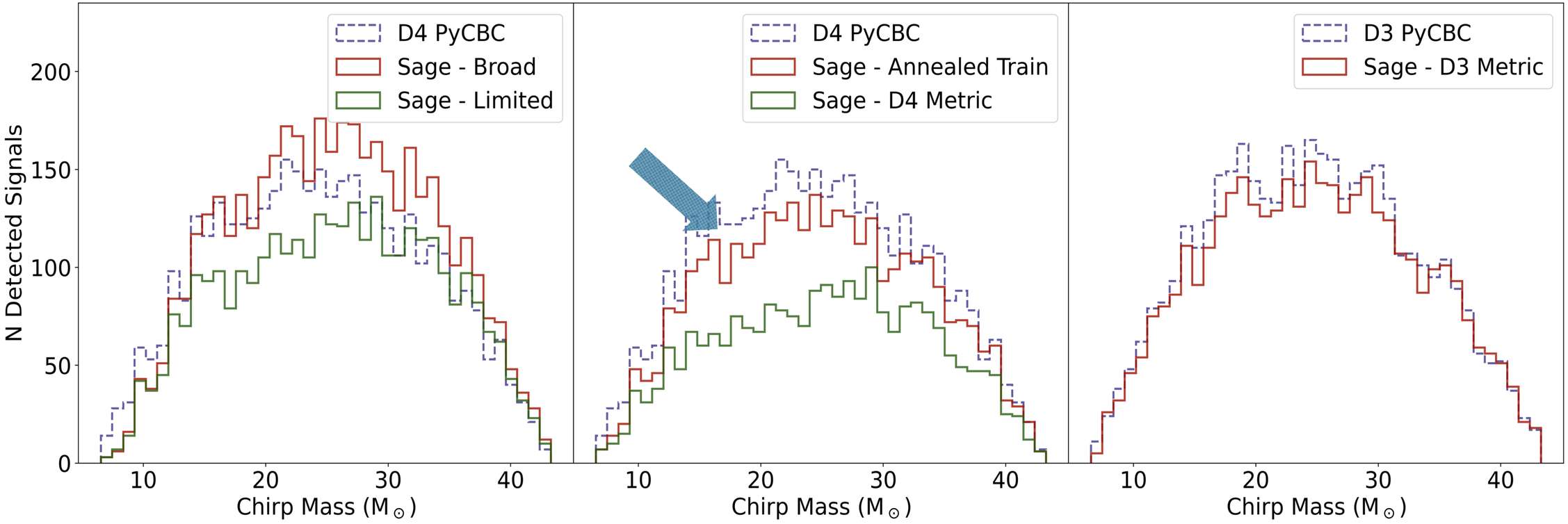
Supplemental Slides

Narenraju Nagarajan
Chris Messenger

arxiv: 2501.13846

IGR

Histograms of chirp mass of the found signals for the injection study at an FAR of 1/month

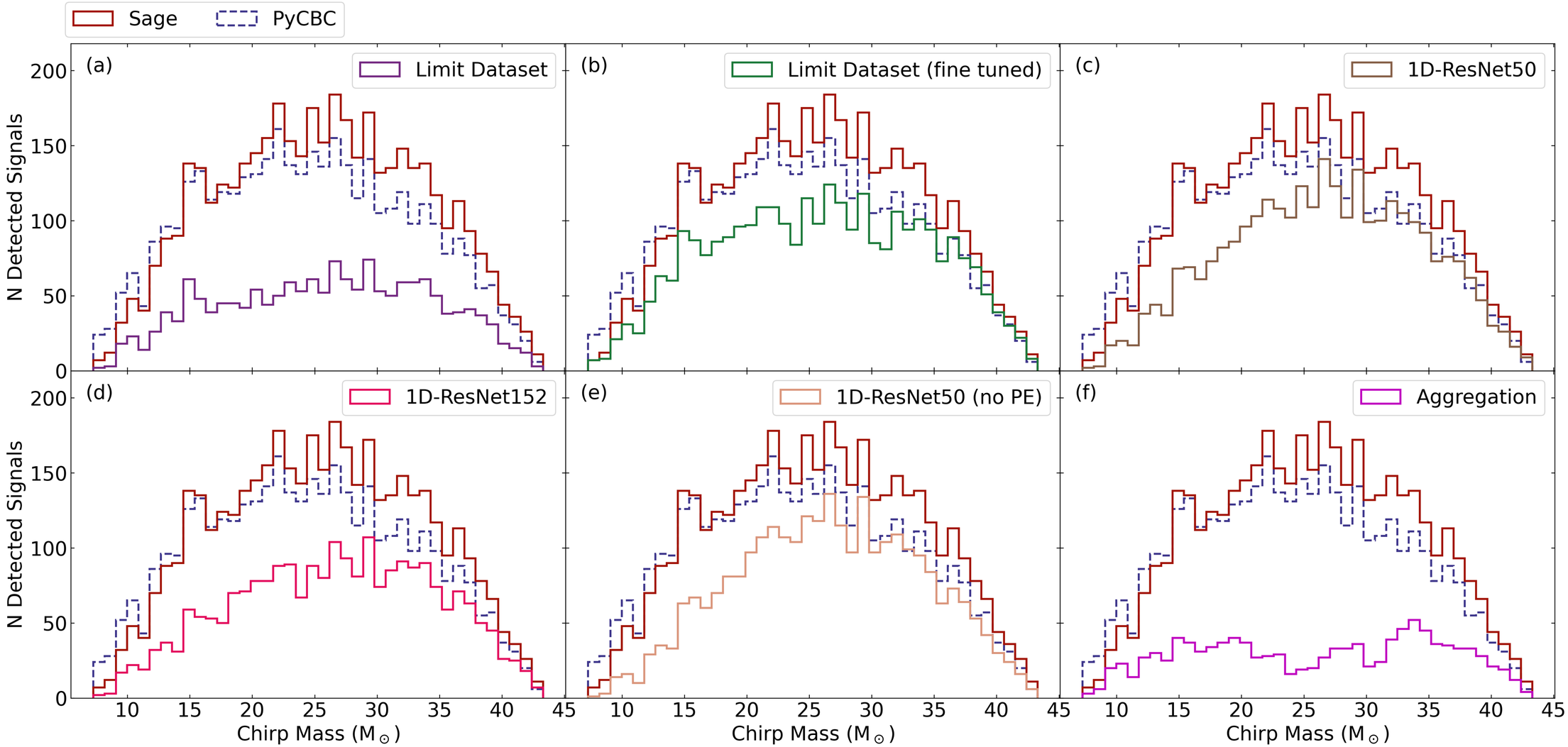


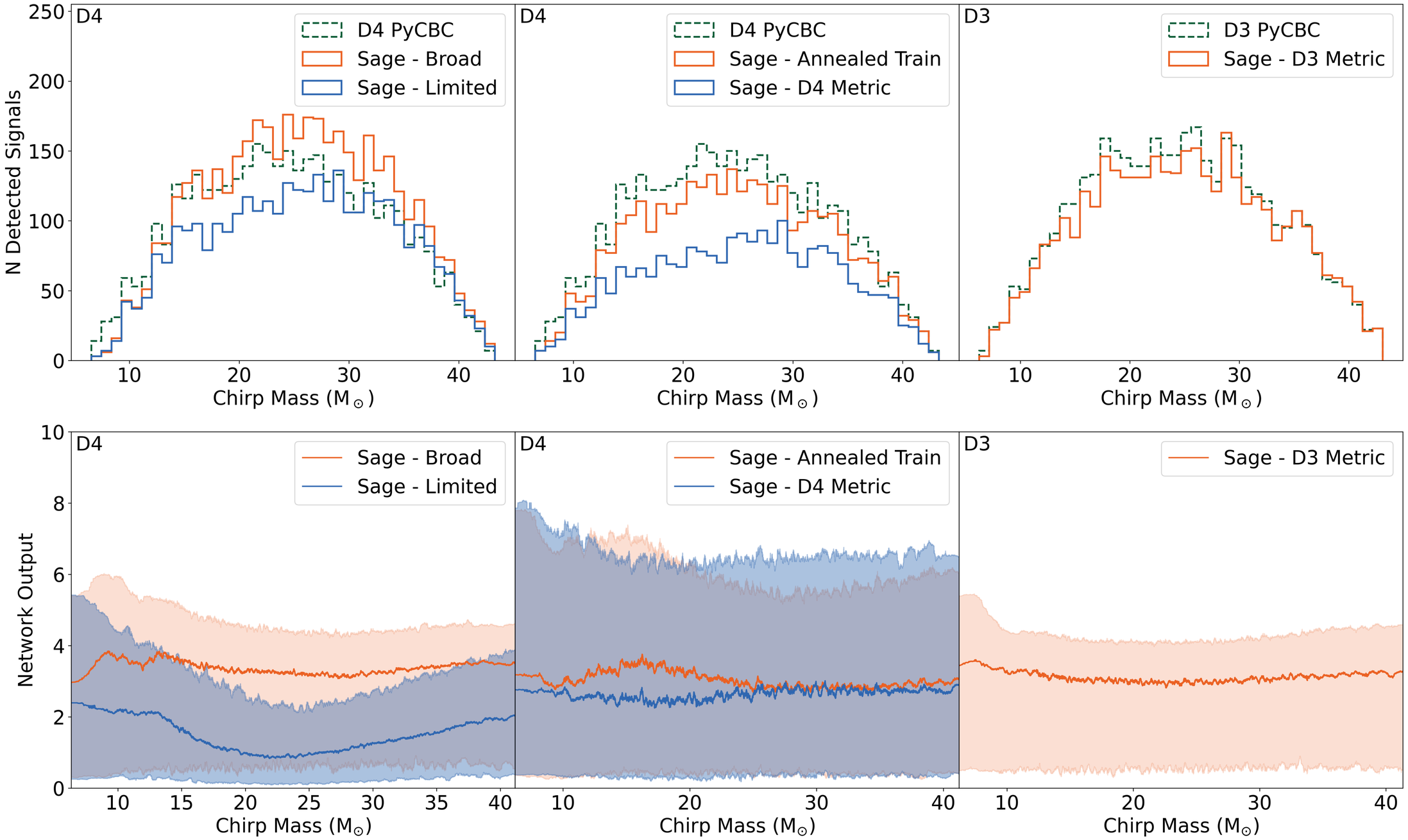
"Sage – Broad" has a broader distribution of noise PSDs than "Sage – Limited".

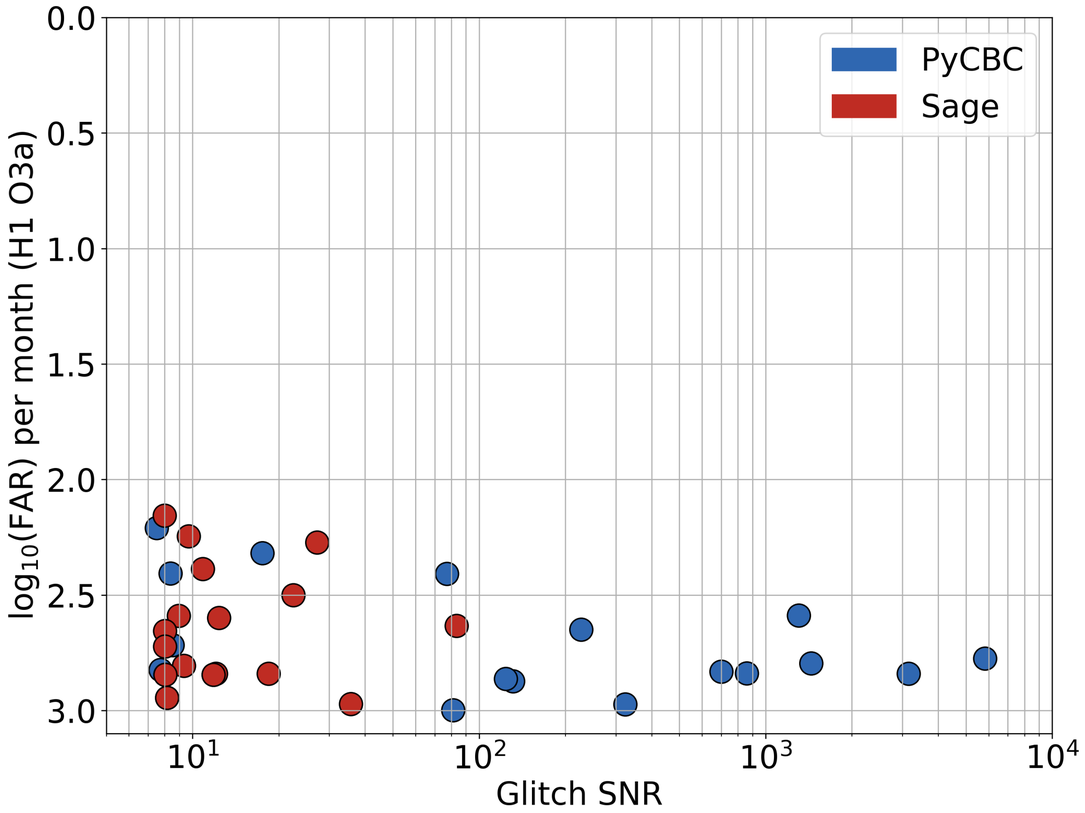
"Sage – D4 Metric" uses the template bank density, $\mathrm{U}(\tau_0, \tau_3)$, to set the mass priors during training. "Sage – Annealed Train" transitions from $\mathrm{U}(\mathrm{m1}, \mathrm{m2})$ to $\mathrm{U}(\tau_0, \tau_3)$ during training.

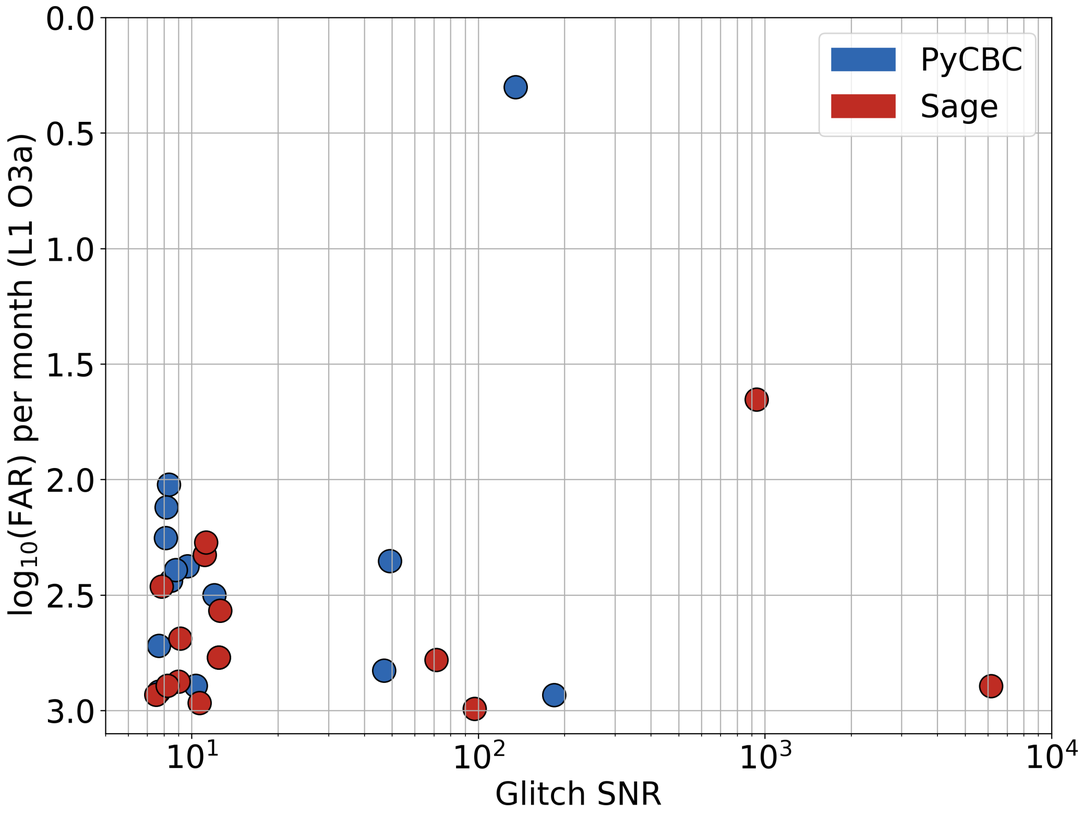
"Sage – D3 Metric" uses $\mathrm{U}(\tau_0, \tau_3)$ to set mass priors for training. Uses simulated coloured Gaussian noise for training and testing.

All aspects of the methodology were kept the same except following changes




D4
D4 PyCBC
Sage - Broad
Sage - Limited
D4
D4 PyCBC
Sage - Annealed Train
Sage - D4 Metric
D3
D3 PyCBC
Sage - D3 Metric
N Detected Signals
Chirp Mass (M⊙)
D4
Sage - Broad
Sage - Limited
D4
Sage - Annealed Train
Sage - D4 Metric
D3
Sage - D3 Metric
Network Output
Chirp Mass (M⊙)



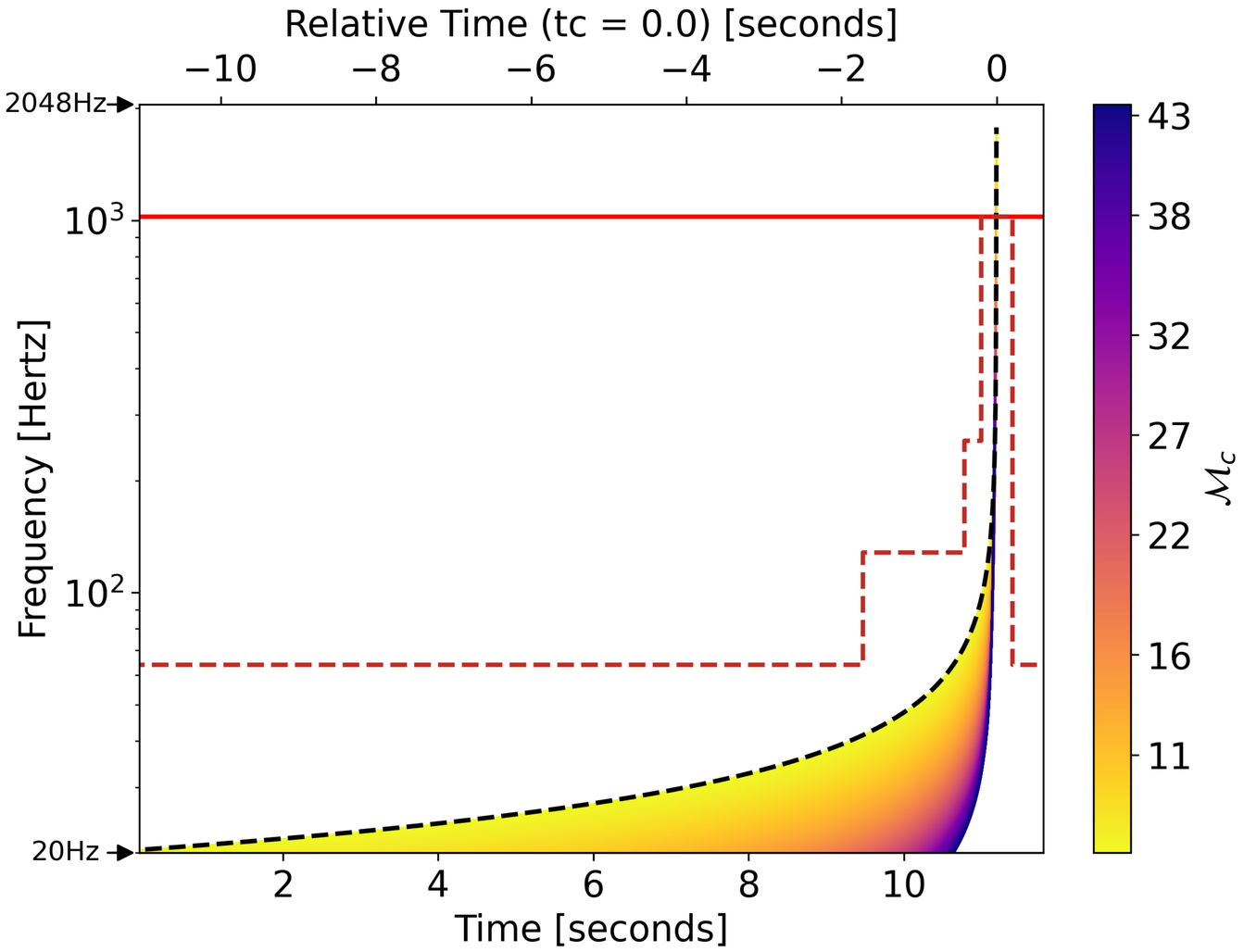


Obtained all triggers by Sage related to glitch events in the O3a testing dataset noise (event times and durations obtained from GravitySpy). Scatter plot shows the ranking statistic and FAR for those triggers as a function of its corresponding glitch SNR for the H1 and L1 detectors. We conclude that Sage can reject glitches effectively in real detector noise.

2

Reduce need for larger receptive field

Compression using multirate sampling



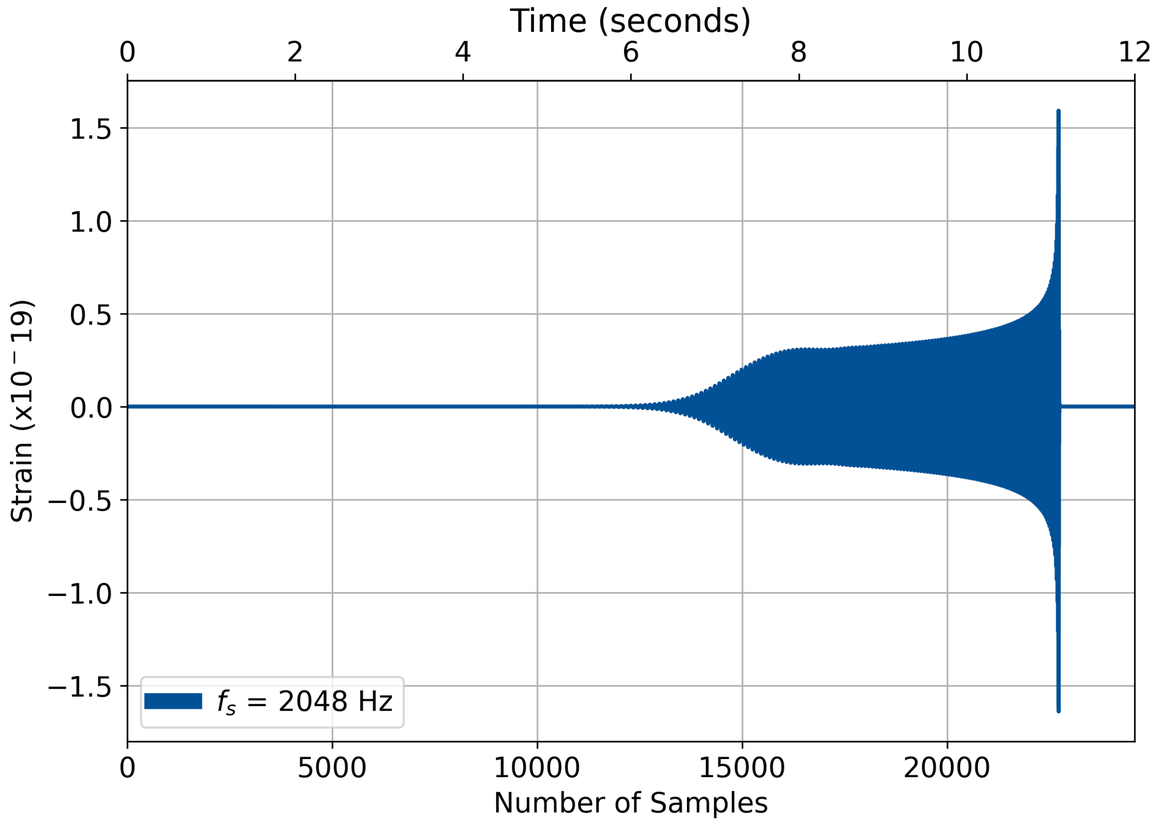
Changing the Nyquist limit from solid red line to dashed red line as a form of time-series compression

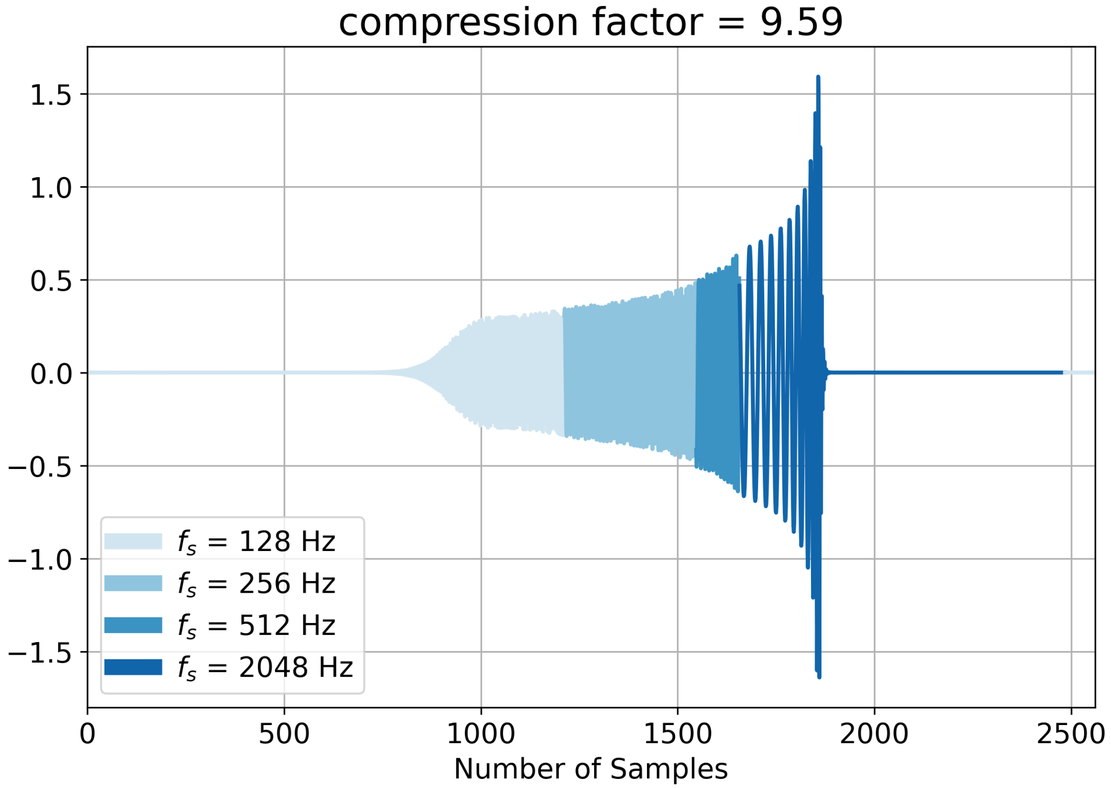
2

Reduce need for larger receptive field

Compression using multirate sampling

Example of compression applied to a BBH gravitational-wave



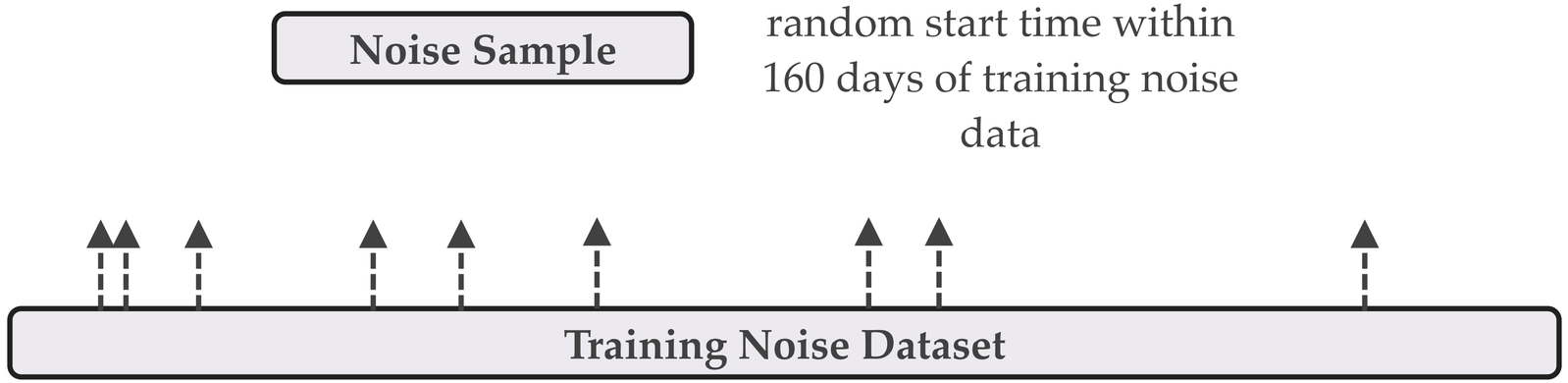


3

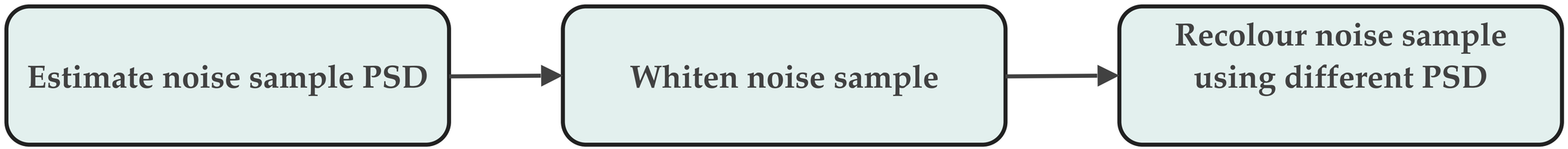
Increased variance provided for noise class via augmentation

Augmentation Method 1

Noise sample has a random start time within 160 days of training noise data

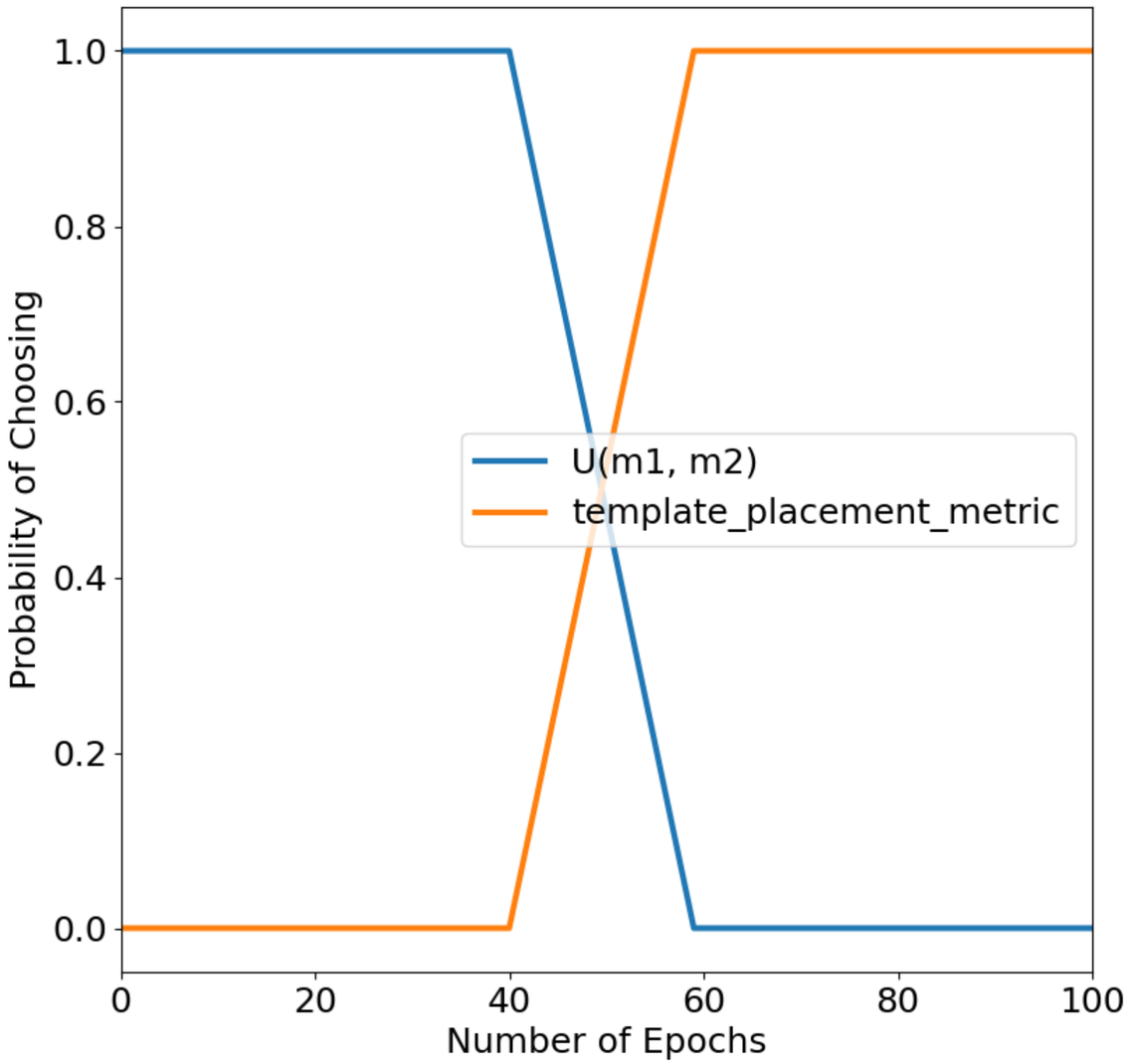


Augmentation Method 2

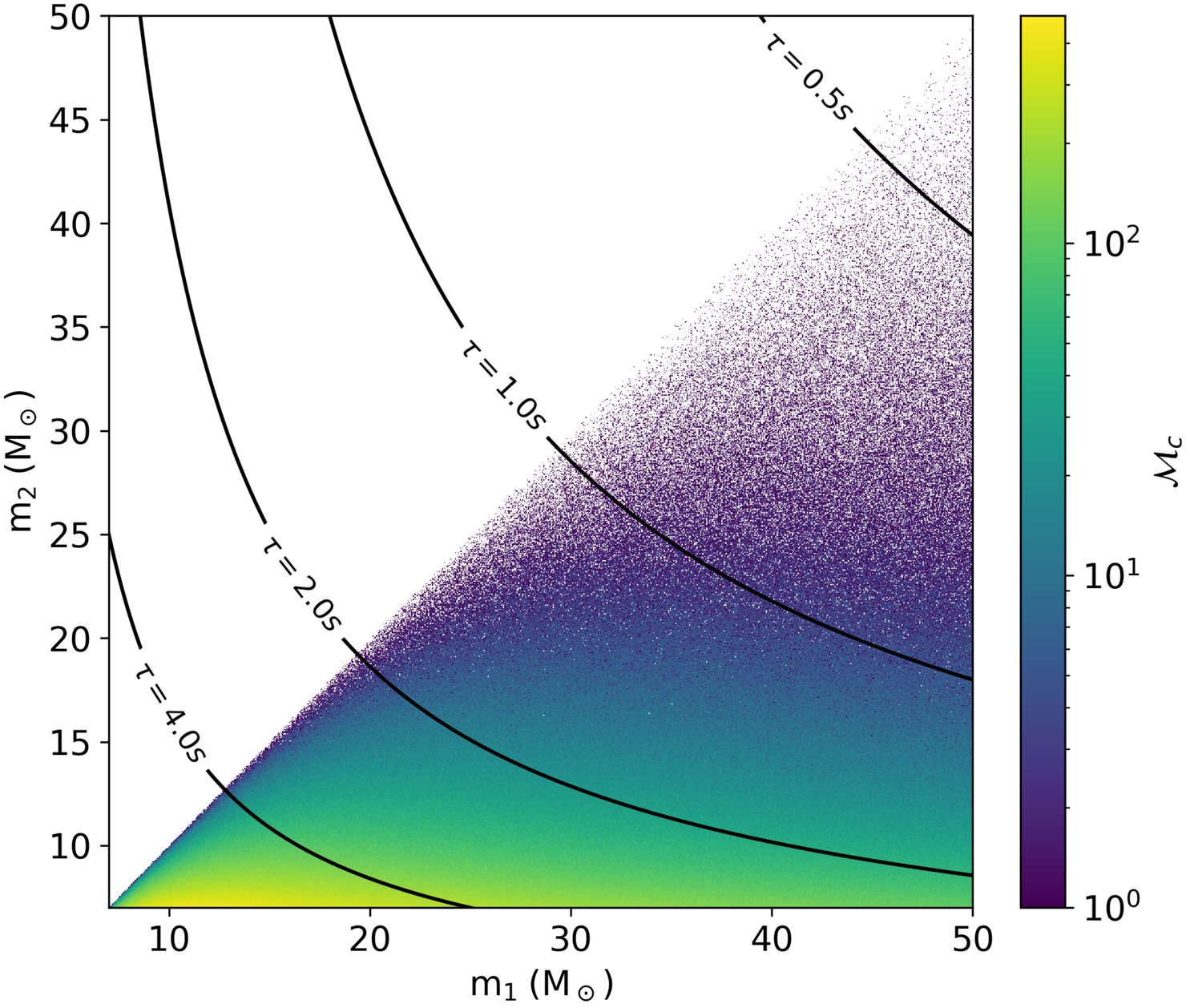


5

Finetuned training strategies to handle biases due to limited samples



Transitioning from U(m1, m2) to $U(\tau_0, \tau_3)$ during training

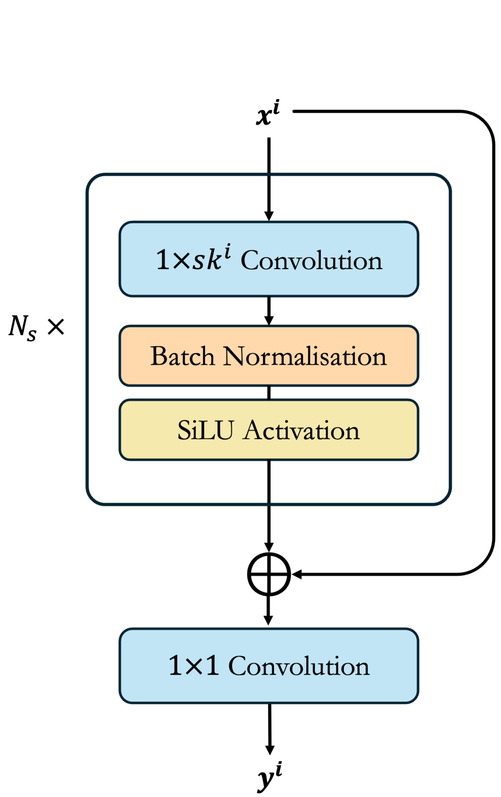


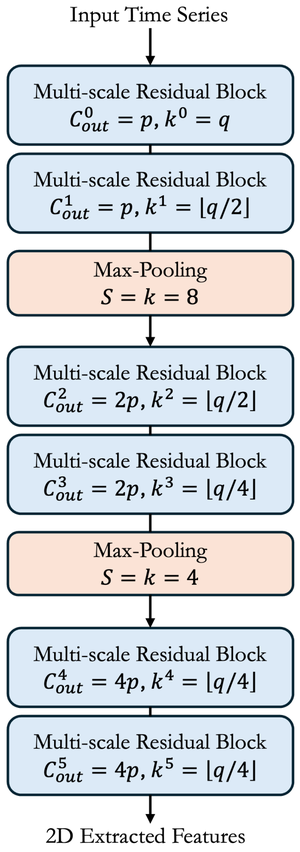
2D Histogram of $U(\tau_0, \tau_3)$ plotted in the mass1 vs mass2 space

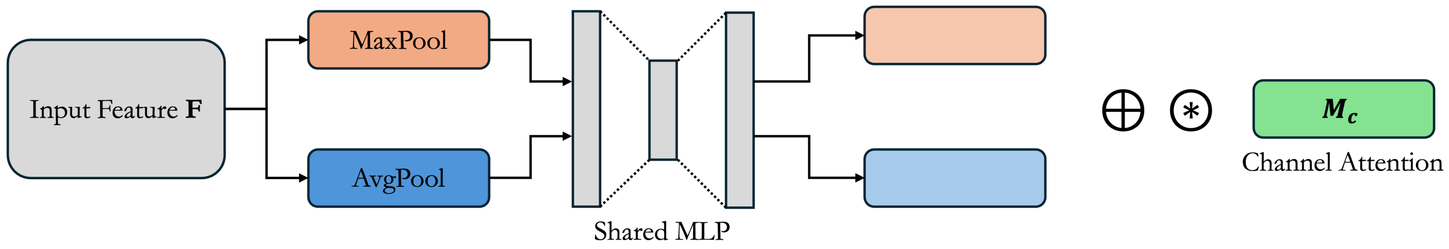
6

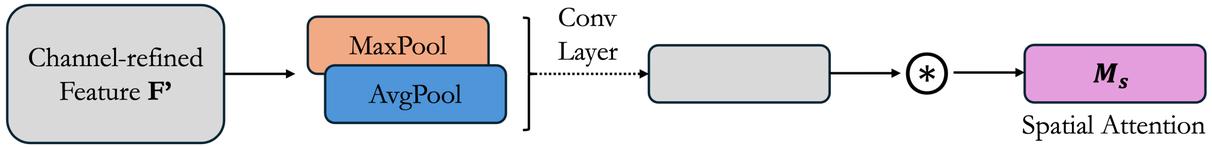
Proper inductive bias used for two stage network architecture

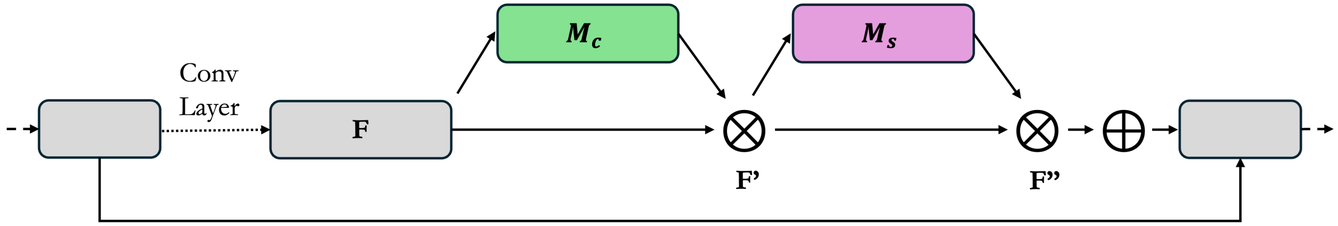
Frontend multiscale feature extractor



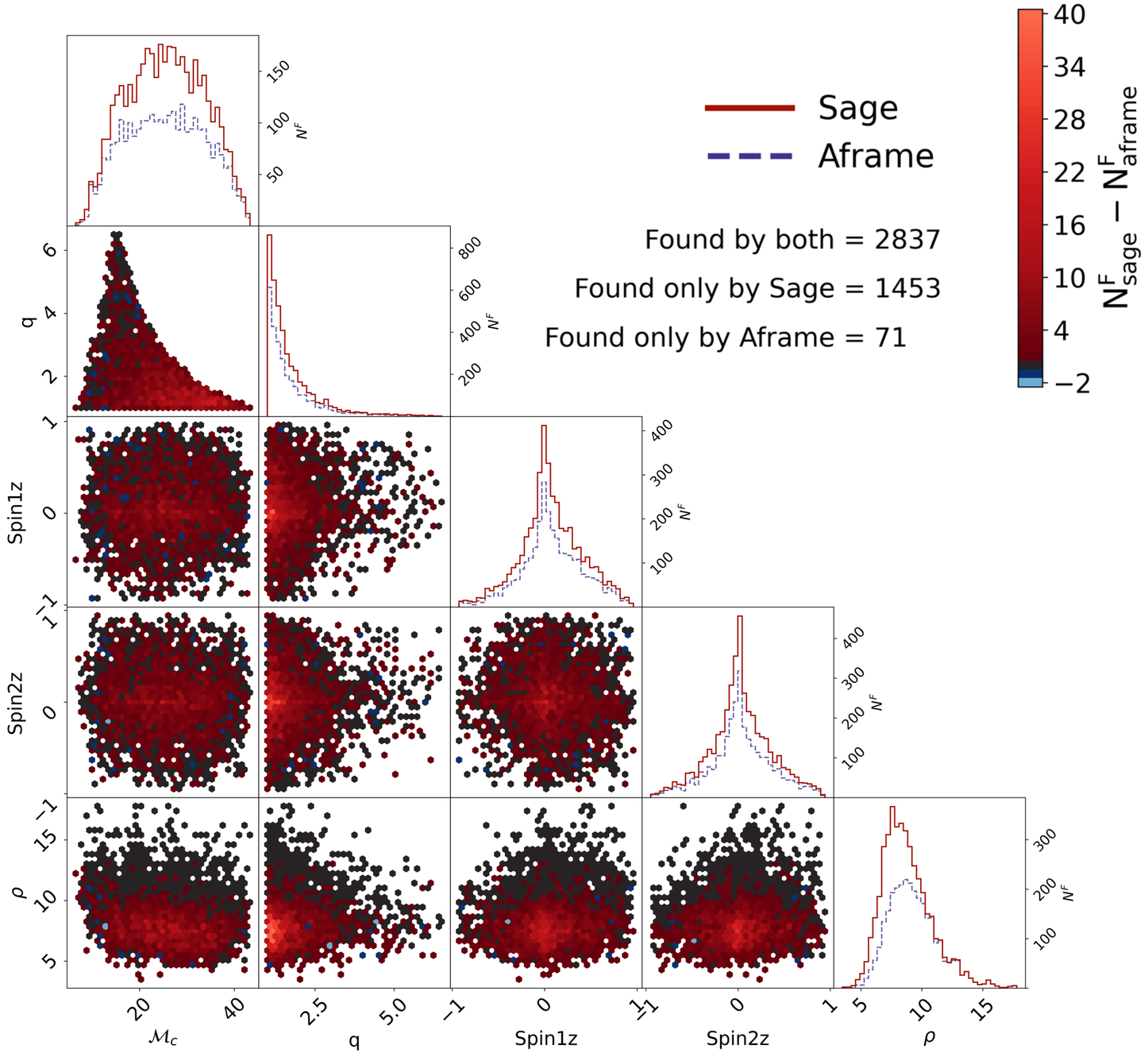




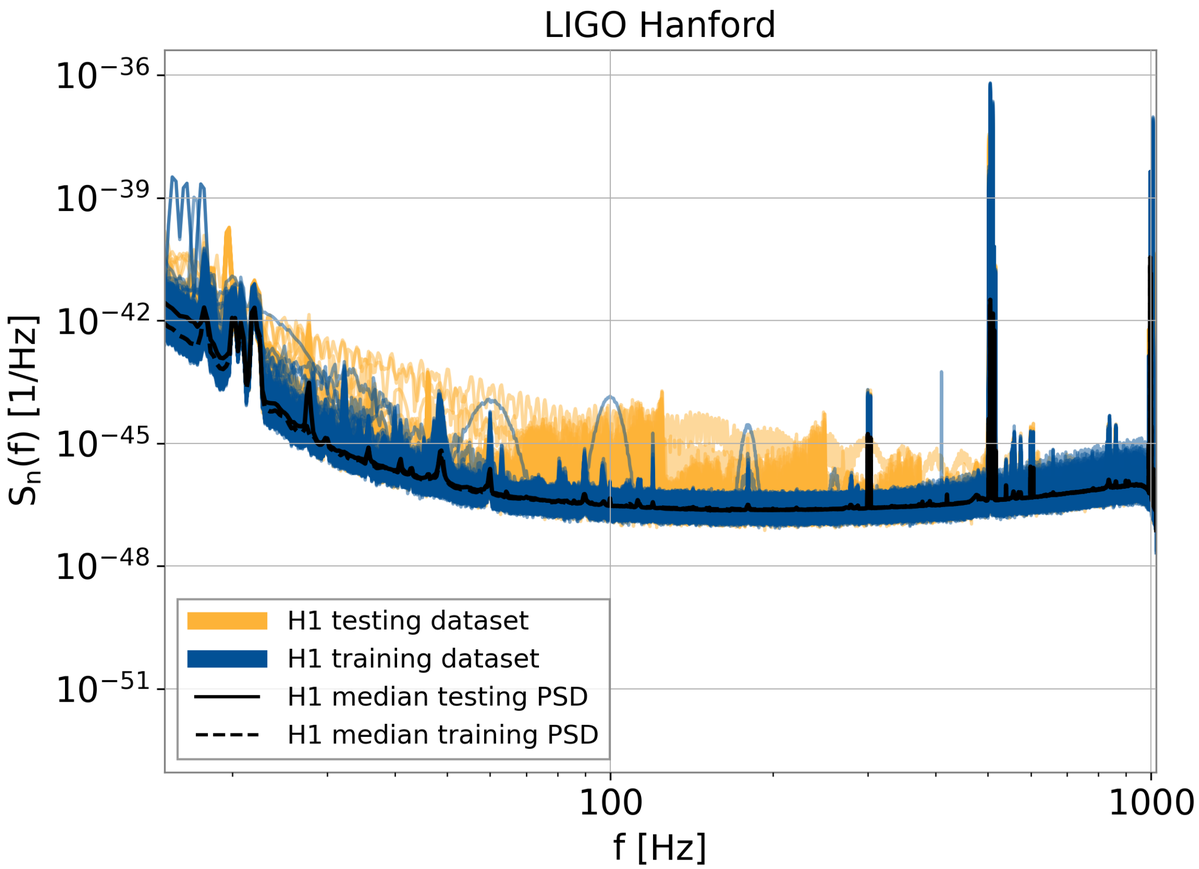


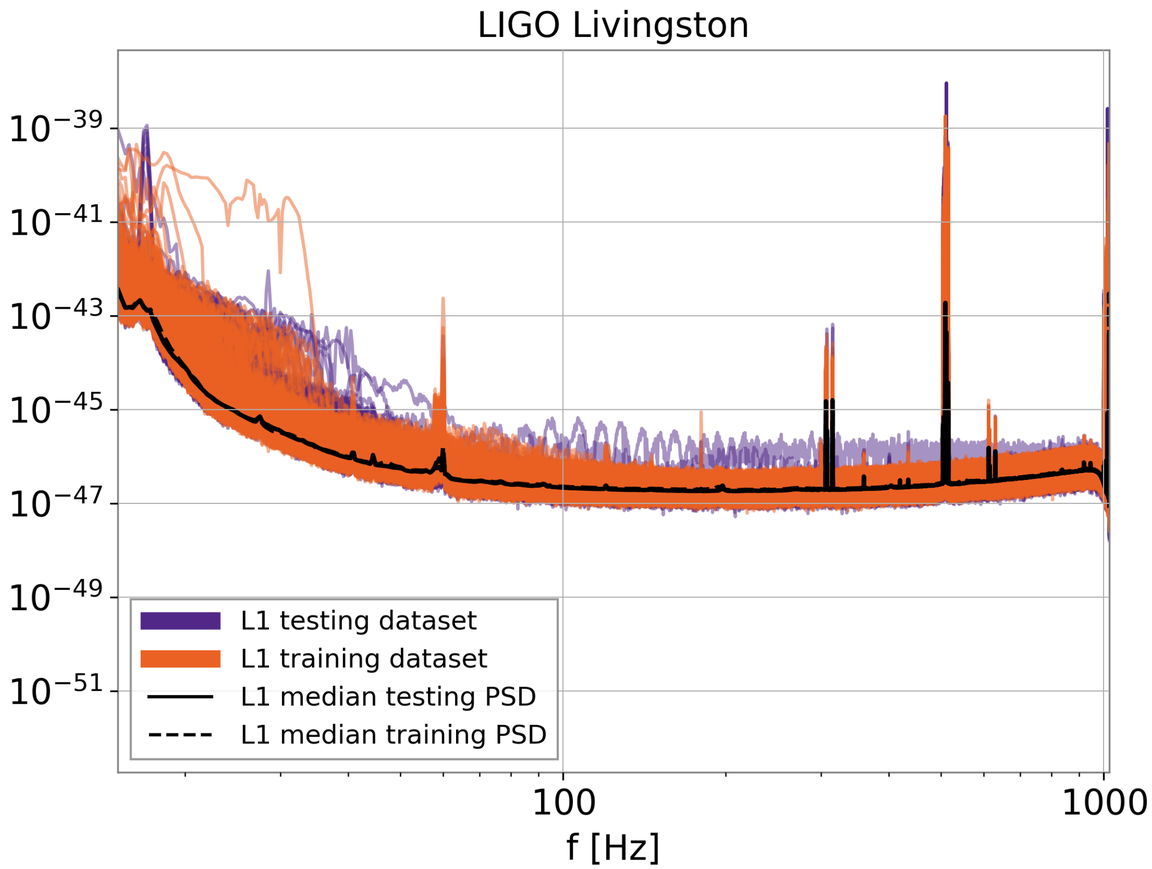


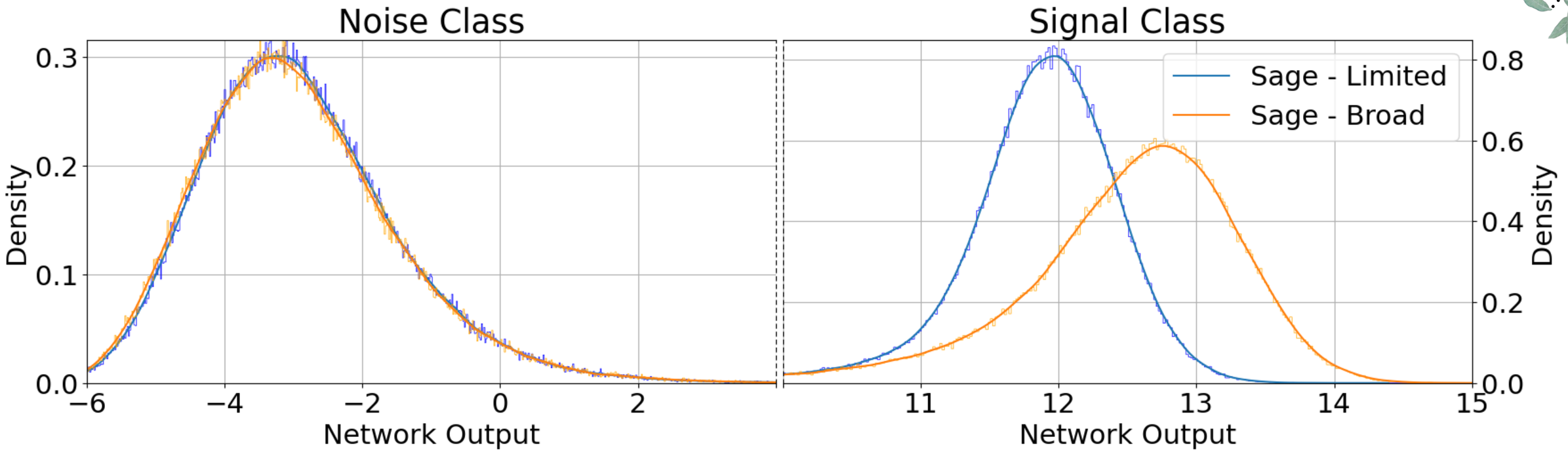
Backend classifier powered by spatial and channel-wise attention



*Aframe is not optimised for the D4 mass distribution in MLGWSC-1. They operate on a broader mass prior [5, 100] Msun.

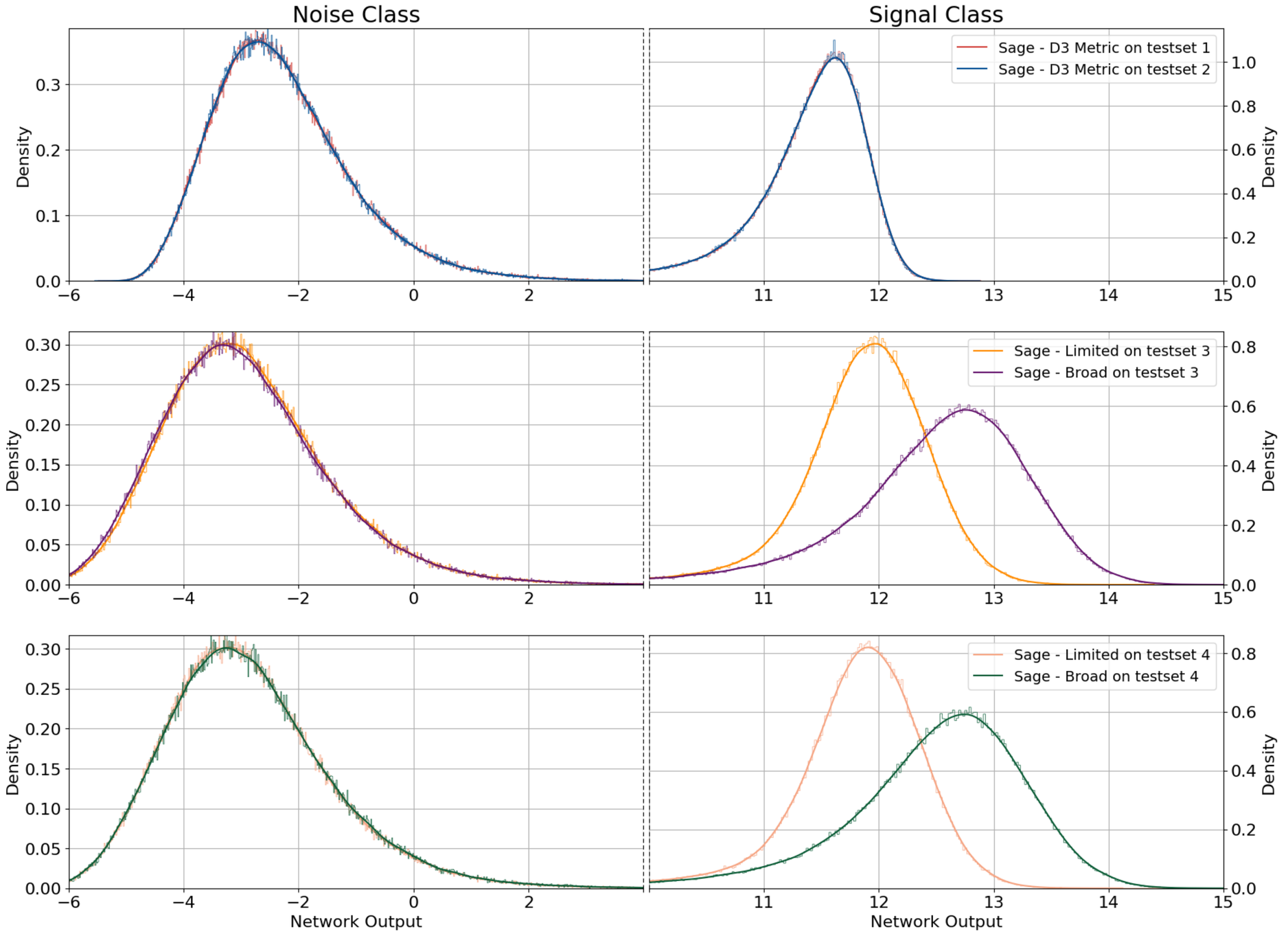

LIGO Hanford
10^{-36}
10^{-39}
10^{-42}
10^{-45}
10^{-48}
10^{-51}
$S_n(f)$ [1/Hz]
H1 testing dataset
H1 training dataset
H1 median testing PSD
H1 median training PSD
100
1000
f [Hz]

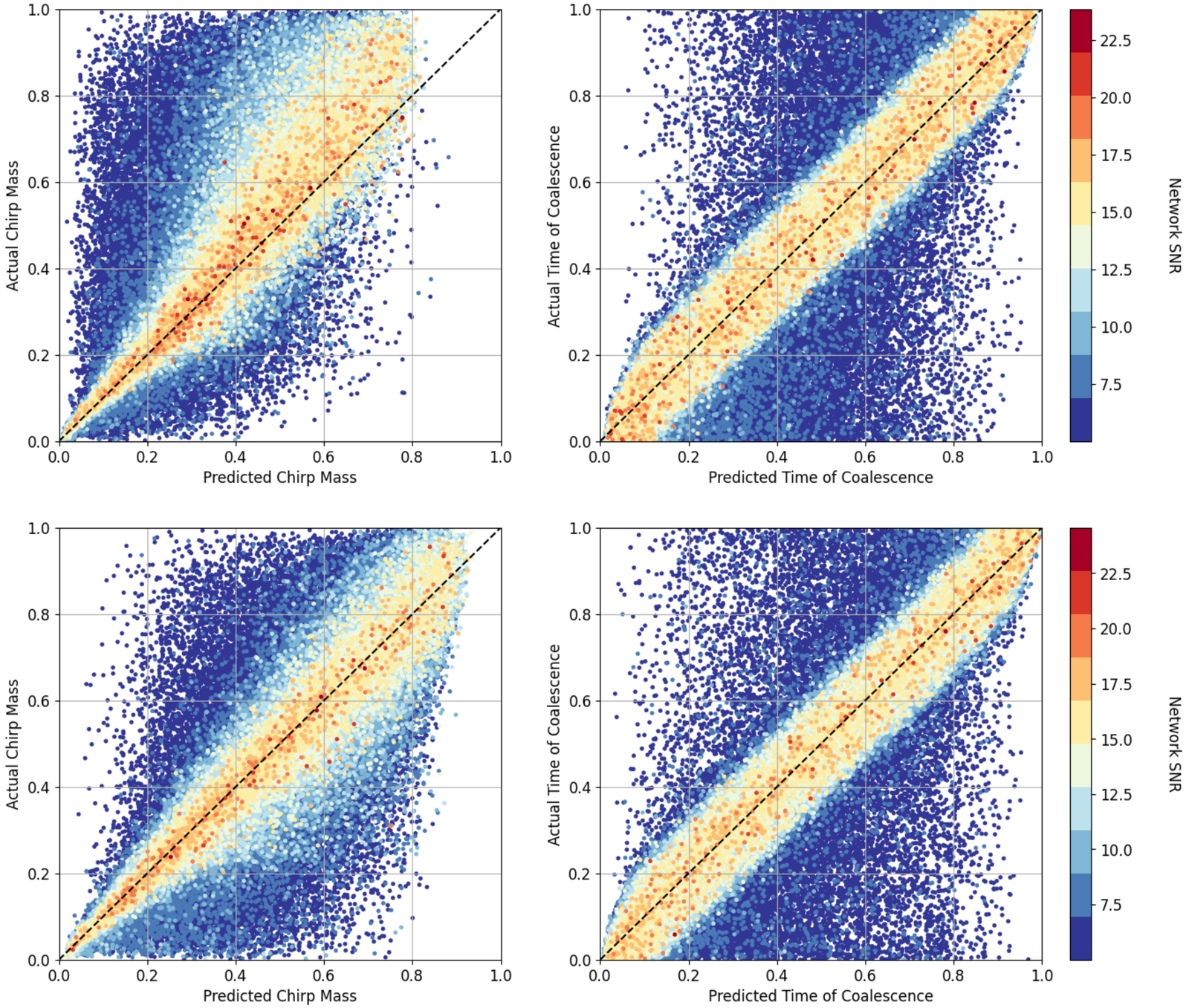

LIGO Livingston
10^{-39}
10^{-41}
10^{-43}
10^{-45}
10^{-47}
10^{-49}
10^{-51}
L1 testing dataset
L1 training dataset
L1 median testing PSD
L1 median training PSD
100
1000
f [Hz]



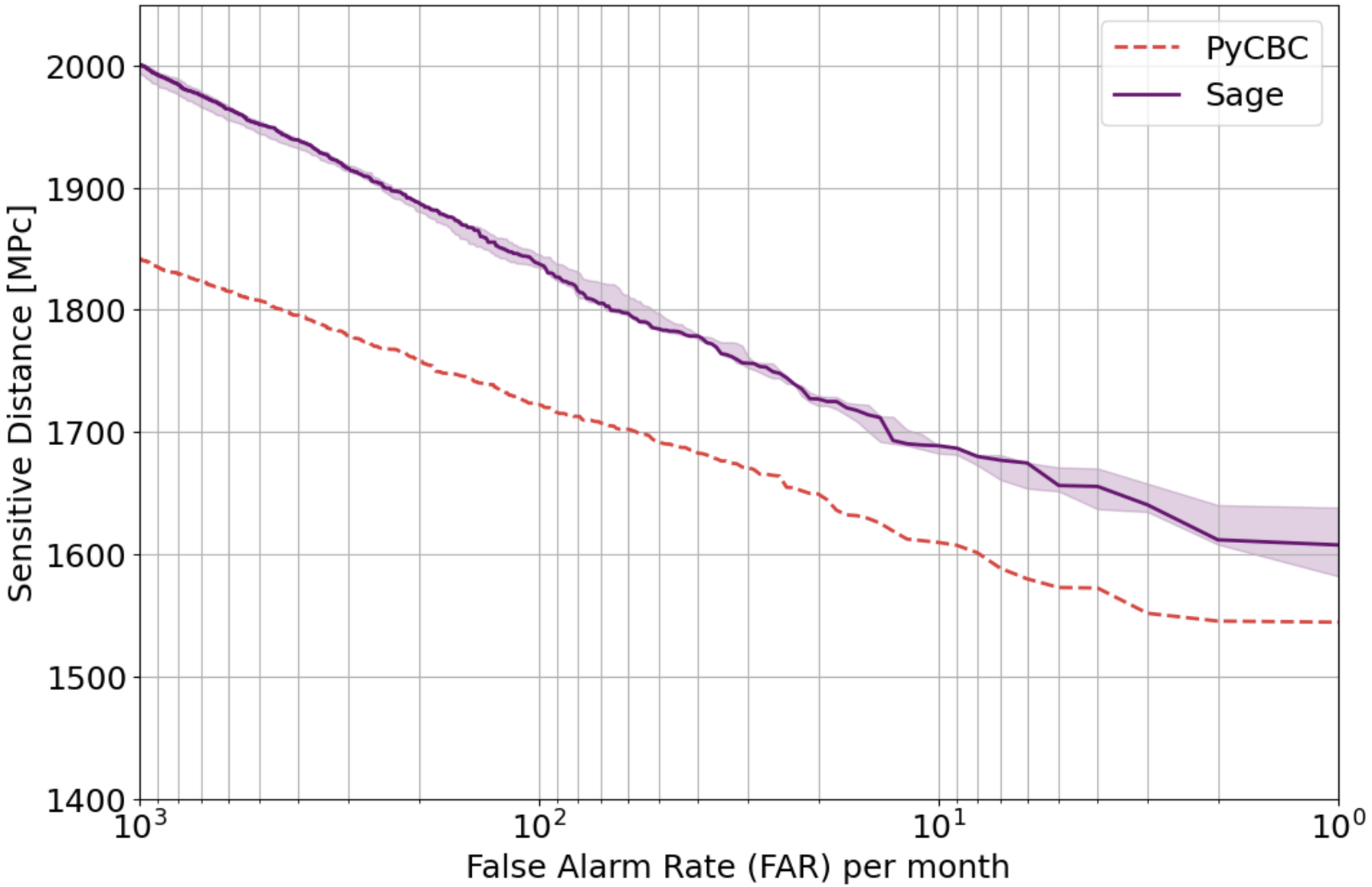
Generated a noise class of 250,000 noise realisations from the O3a noise PSDs. Generated a signal class dataset by injecting the same signal into these noise realisations.

Comparing the normalised histogram of network outputs for “Sage – Broad” and “Sage – Limited” suggests that training a model with a larger variance in noise PSDs allows it to be more confident about a given signal compared to a model that is not.

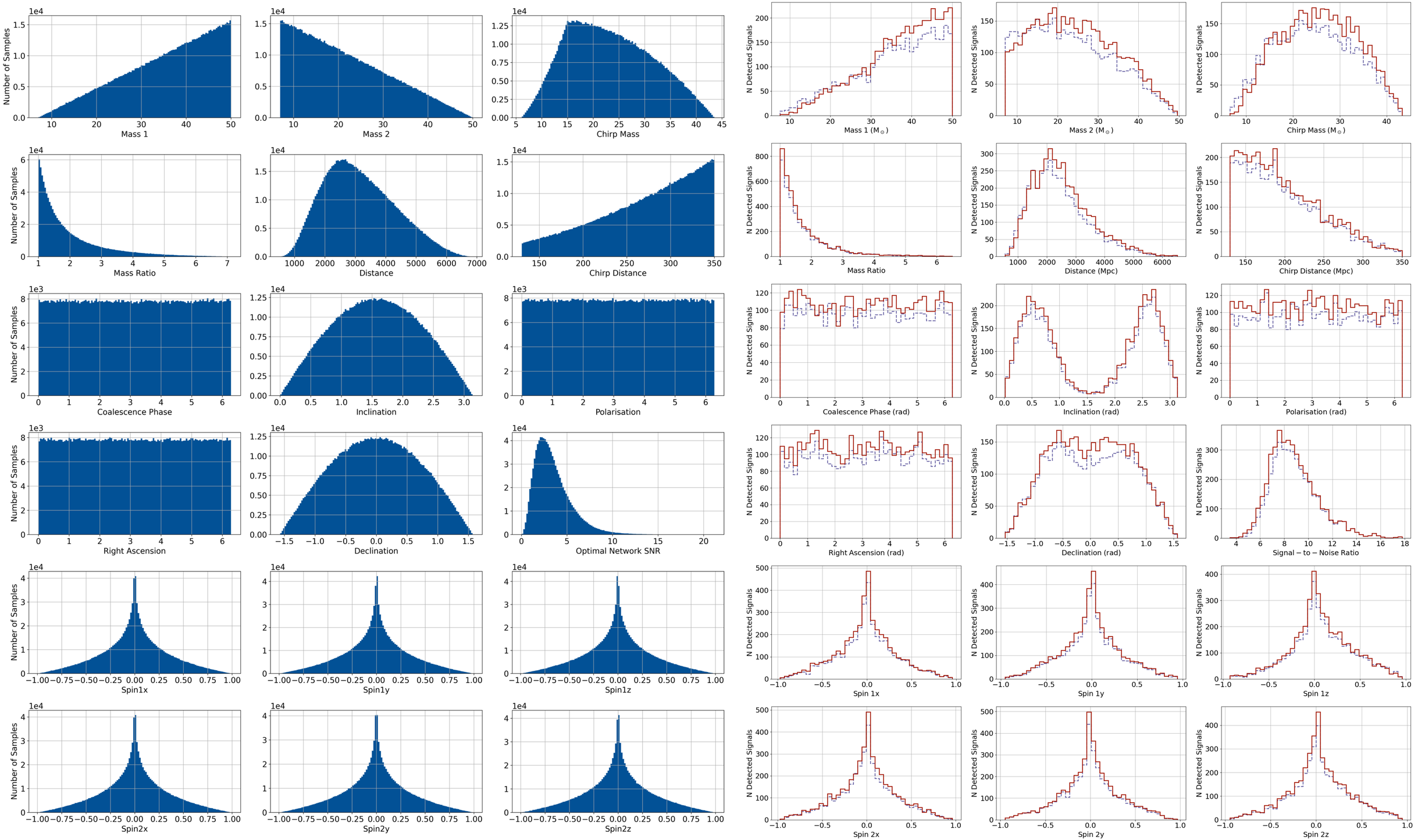

Noise Class
Signal Class
Sage - D3 Metric on testset 1
Sage - D3 Metric on testset 2
Sage - Limited on testset 3
Sage - Broad on testset 3
Sage - Limited on testset 4
Sage - Broad on testset 4
Density
Network Output



Actual value of GW parameters: chirp mass and time of coalescence, plotted against the network predicted value for a validation epoch in different Sage runs. The top row corresponds to the Sage - Annealed Train run and the bottom row to the Sage - Broad run.



Error bar obtained using 3 runs of Sage with different training seeds.



PyCBC - dotted blue, Sage – solid red lines