

Learning a Robust Foreground Models for Multi-Source GW Inference

One learned foreground model for calibrated noise reconstruction and foreground-aware source inference

He Wang

ICTP-AP, University of Chinese Academy of Sciences (UCAS)

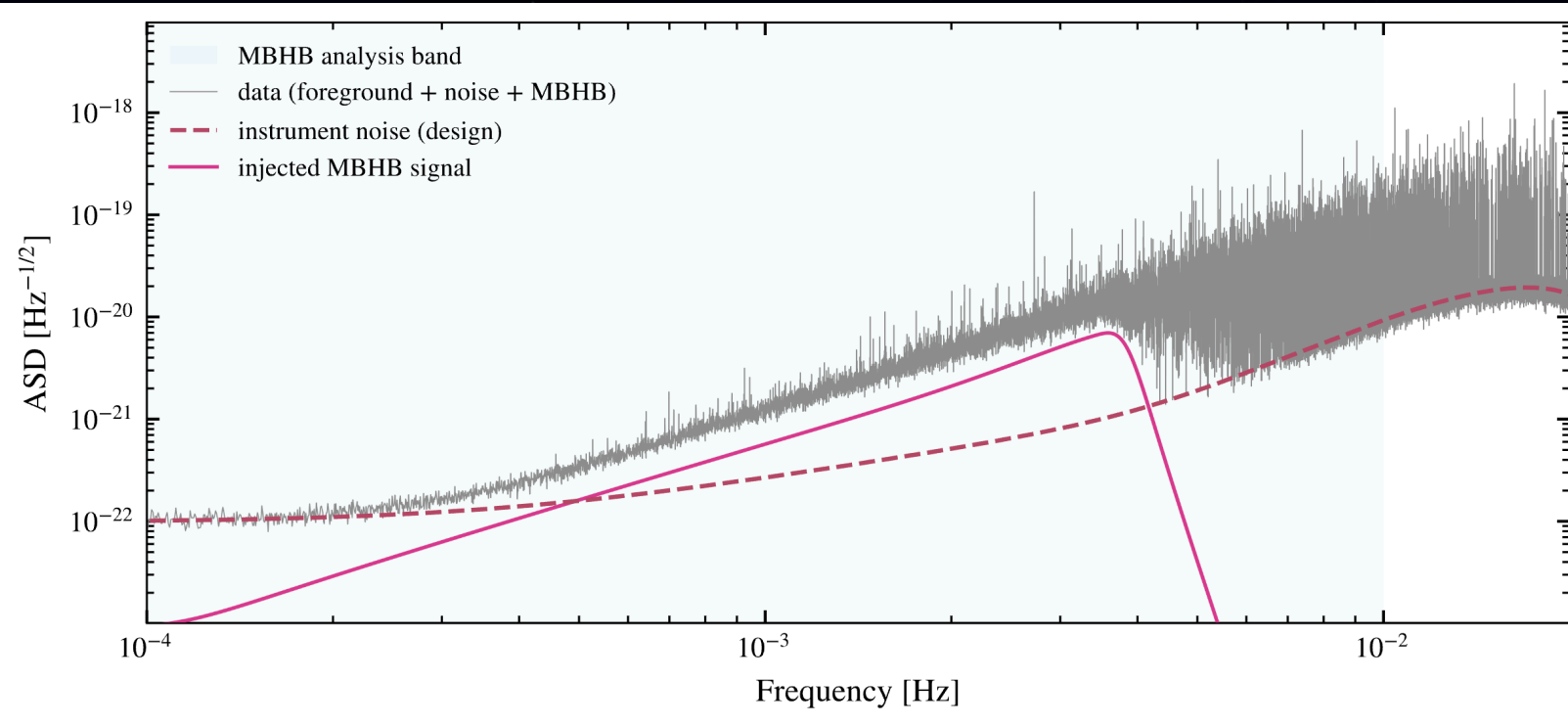
Taiji Laboratory for Gravitational Wave Universe, UCAS

*In collaboration with **Dingkai Zhang**, South China Normal University*

(in prep.)

The millihertz data stream is signal-dominated, not noise-dominated.

Tens of millions of Galactic binaries overwhelm the instrumental floor across the band where many high-value sources live.



Credit: Dingkai Zhang

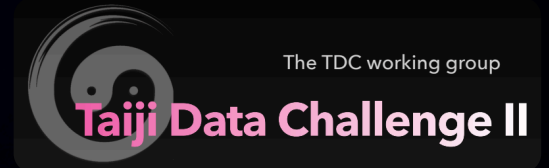
One-year Taiji simulation

- $\sim 4.5 \times 10^7$ Galactic binaries
- numerical Taiji orbit
- unequal-arm TDI-2.0
- frequency-domain A/E channels

Why is this difficult?

- Confusion dominates: $\sim 0.3\text{--}10$ mHz
- No signal-free frequency region
- Foreground and instrumental noise cannot be separated directly

**How can we estimate the instrumental PSD
when no frequency region contains only noise?**



There is no clean off-source measurement of the instrumental PSD.

Signal confusion creates two coupled inference failures

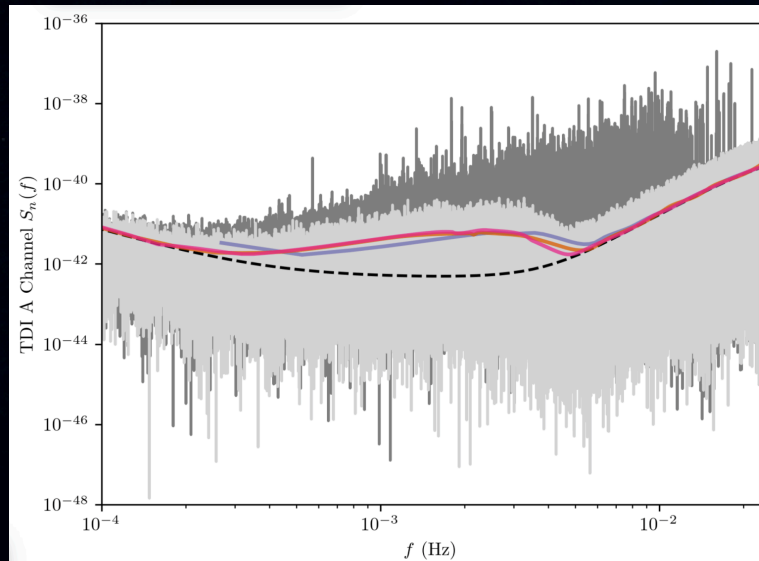
Foreground contamination biases the noise model and leaves structured residuals that degrade downstream source inference.

Noise characterization fails

$$\hat{S}_{\text{res}}(f) \neq S_{\text{inst}}(f)$$

- foreground power contaminates the residual periodogram
- the inferred PSD is systematically distorted
- frequency-dependent likelihood weights become incorrect

Galactic foreground biases instrumental PSD estimation.



Littenberg & Cornish (2023)

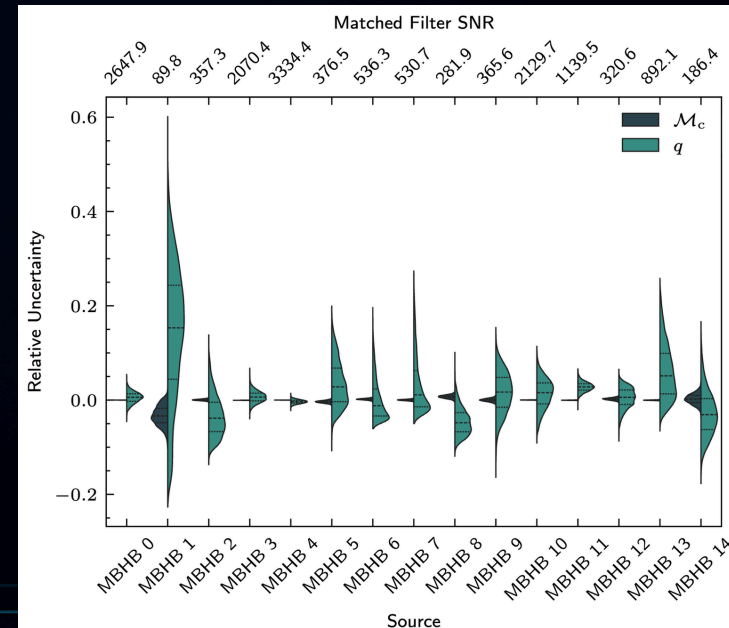
Target-source model is misspecified

$$r(f; \theta) = d(f) - h(f; \theta) = g_{\text{GB}}(f) + n(f)$$

$$r(f; \theta) \not\sim \mathcal{N}(0, S_{\text{inst}}(f))$$

- the residual contains structured foreground power
- the assumed Gaussian-noise likelihood is misspecified
- posteriors can broaden, shift, or become overconfident

Foreground residuals degrade source inference.

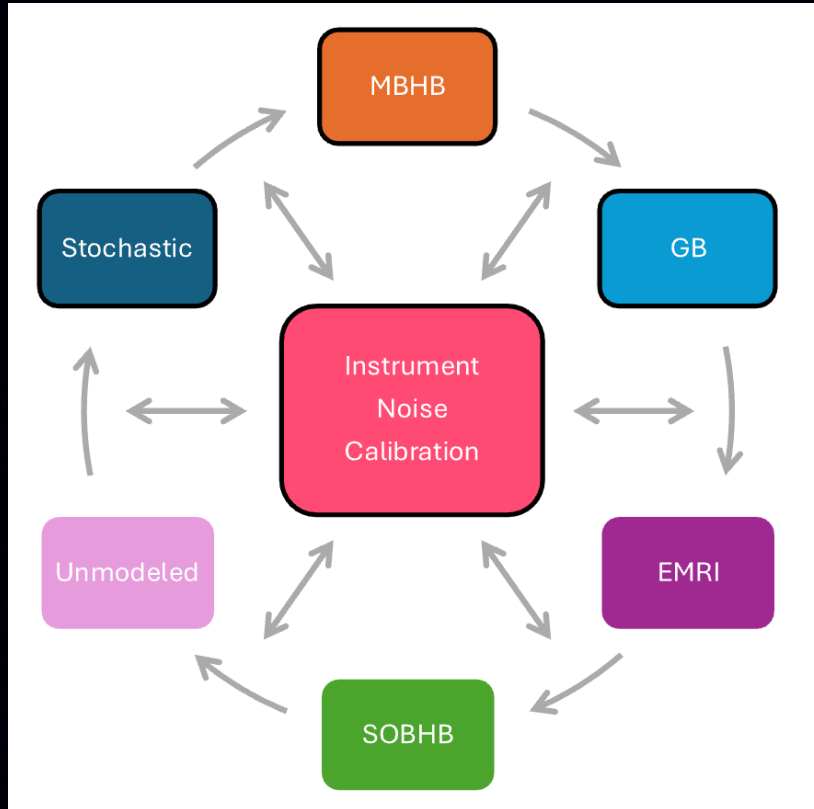


Deng, Babak, and Marsat (2025)

A contaminated residual means both a distorted PSD and a misspecified source likelihood.

The global fit solves the problem—but at a price

A global fit resolves the coupling in principle, but does so by explicitly modeling individual sources.



Katz, Karnesis, Korsakova, Gair, and Stergioulas (2025)

What the global fit provides

- individual source catalogs
- physical noise posterior
- joint Bayesian consistency
- interpretable residuals

What it costs

- trans-dimensional sampling
- repeated waveform generation
- blocked iterative convergence
- computational cost scales with source number

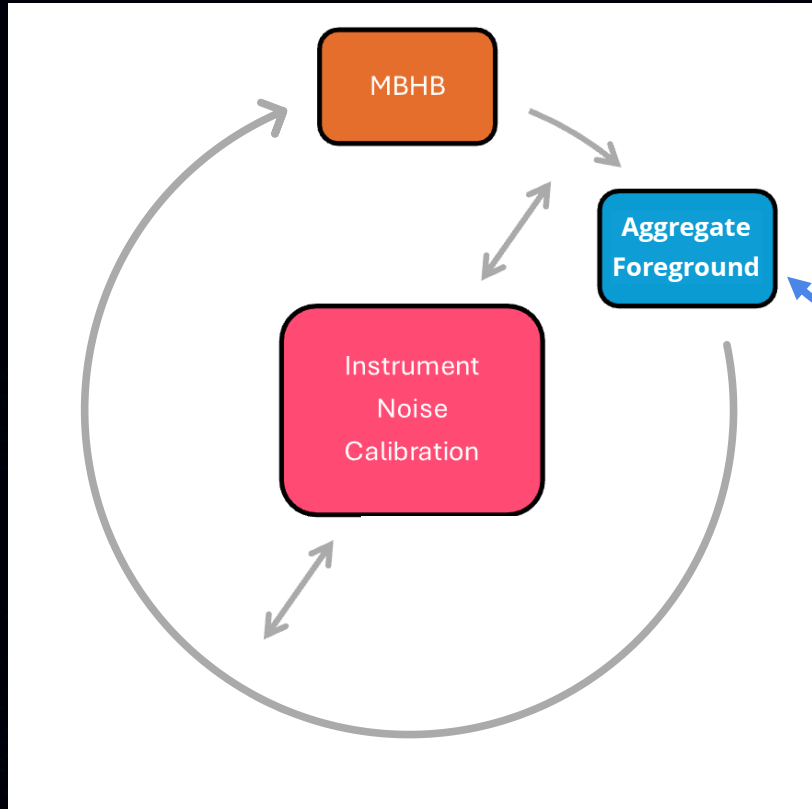
Scientific Question

Do we need the **identity of every Galactic binary** to infer the instrumental noise or validate one target source?

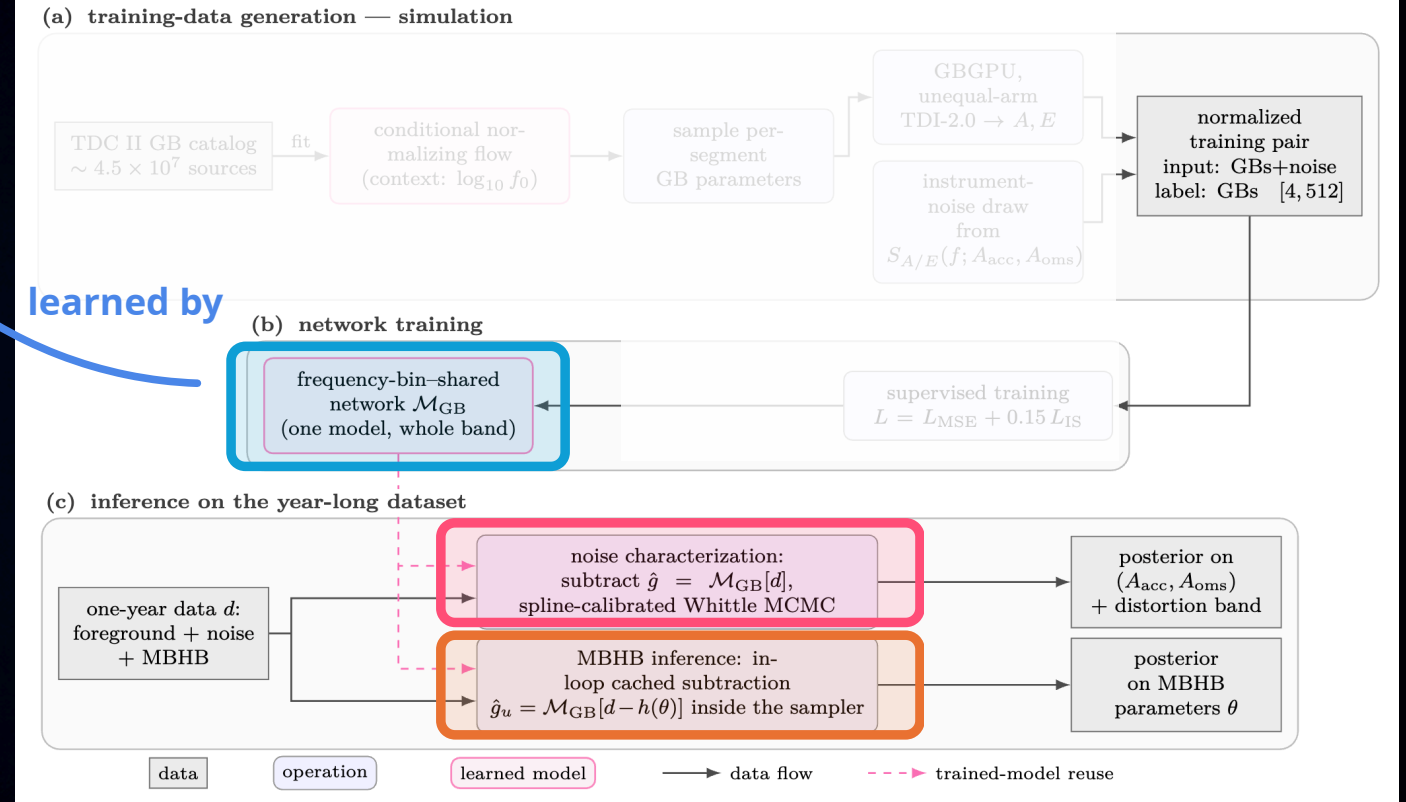
For noise calibration and target inference, the aggregate foreground may be more important than the identity of every source.

Our key design choice: learn the aggregate foreground

The network predicts the total Galactic foreground, which is then reused for noise characterization and source inference.



Katz, Karnesis, Korsakova, Gair, and Stergioulas (2025)
(adapted)



Pipeline overview

**No catalog in, no catalog out.
Aggregate foreground is the prediction target.**

The network learns a population—not a catalog

Every training sample is a new foreground realization, preventing catalog memorization.

Population-conditioned simulation

TDC II catalog

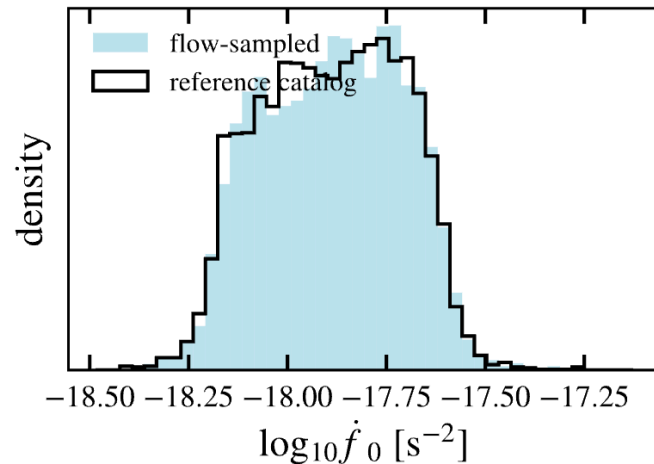
- sample source parameters and counts
- GBGPU + Taiji response
- TDI-2.0 A/E segment

Conditional normalizing flow:

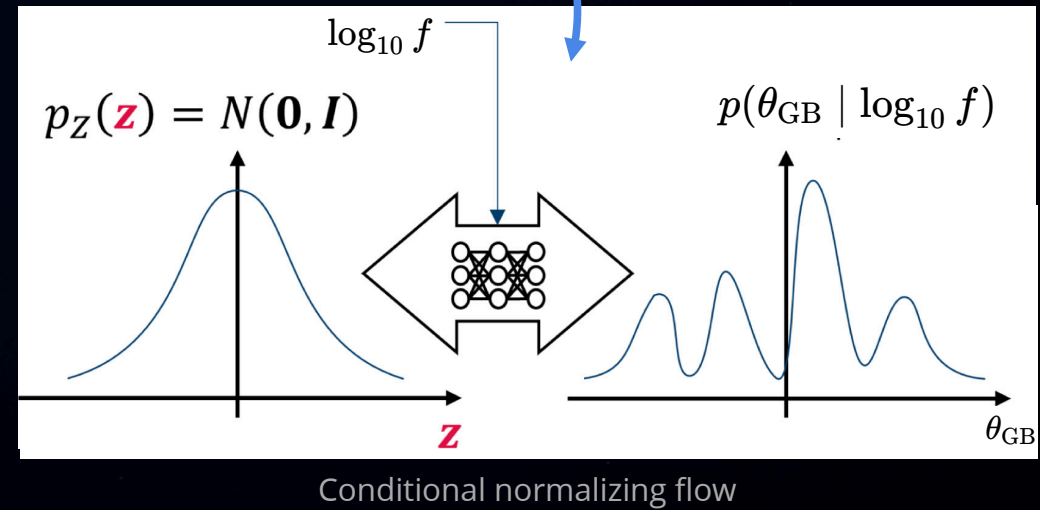
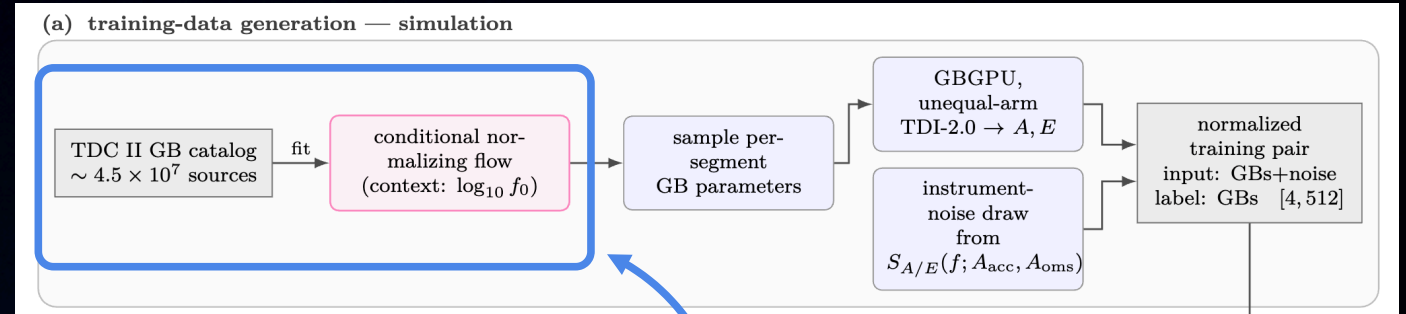
- **Target:** $p(\theta_{\text{GB}} \mid \log_{10} f)$

Key choices:

- **full catalog, no SNR split**



Population validation near 1 mHz



No realization is ever repeated.

One shared network covers the full analysis band

Local spectral morphology is sufficiently universal for one shared network to operate across the entire band.

One frequency-bin-shared network

Input representation

$$\mathbf{x} \in \mathbb{R}^{4 \times 512}$$

Local 512-bin window in A/E channels

Why one shared network?

- local morphology is translation-invariant
- dense and sparse regions share the same representation
- frequency enters only through conditioning

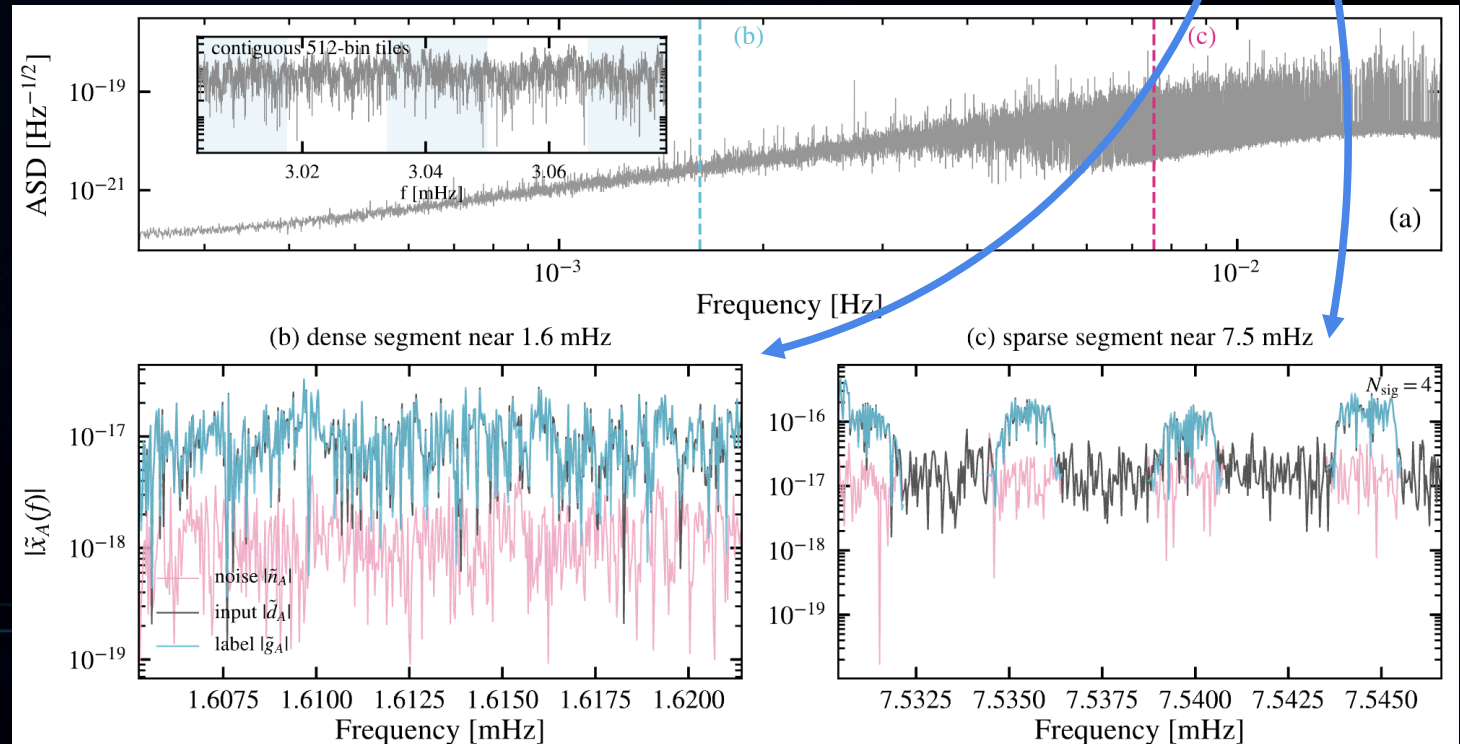
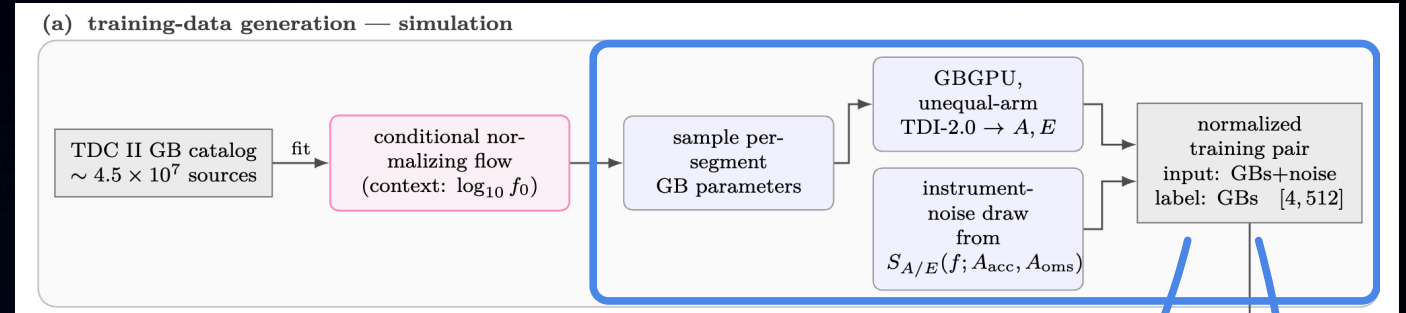
Sliding inference

- overlapping windows
- crop borders
- stitch predictions

Key choices:

- random source counts: $N_{\text{ref}}(f) \pm 2\sqrt{N_{\text{ref}}(f)}$
- bleed padding near segment edges

Shared morphology enables a shared network.



The architecture is simple; the training objective makes the residual usable

The loss is designed for the downstream likelihood, not just foreground reconstruction.

Calibrate the residual, not just the prediction.

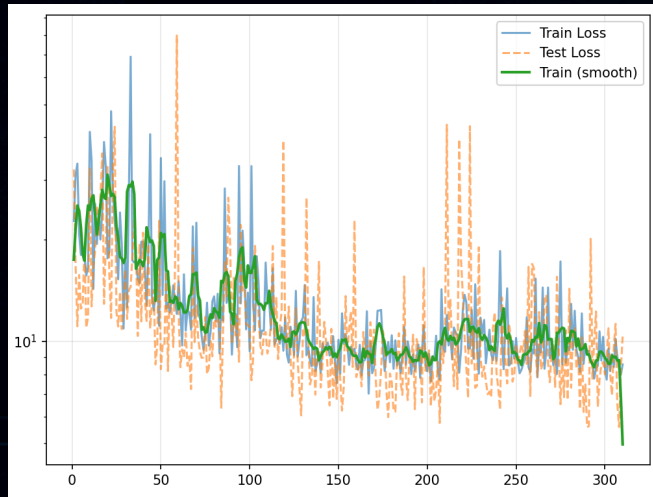
Downstream-aware objective

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda \mathcal{L}_{\text{IS}}, \quad \lambda = 0.15$$

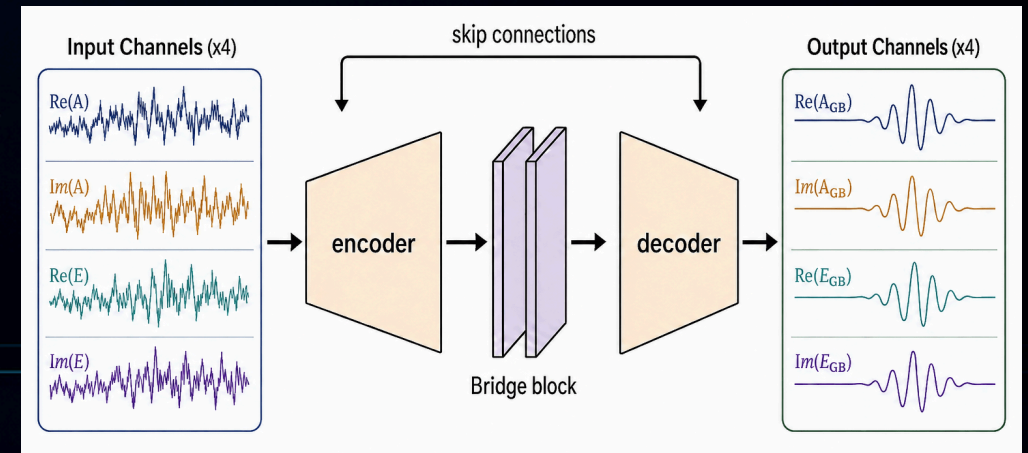
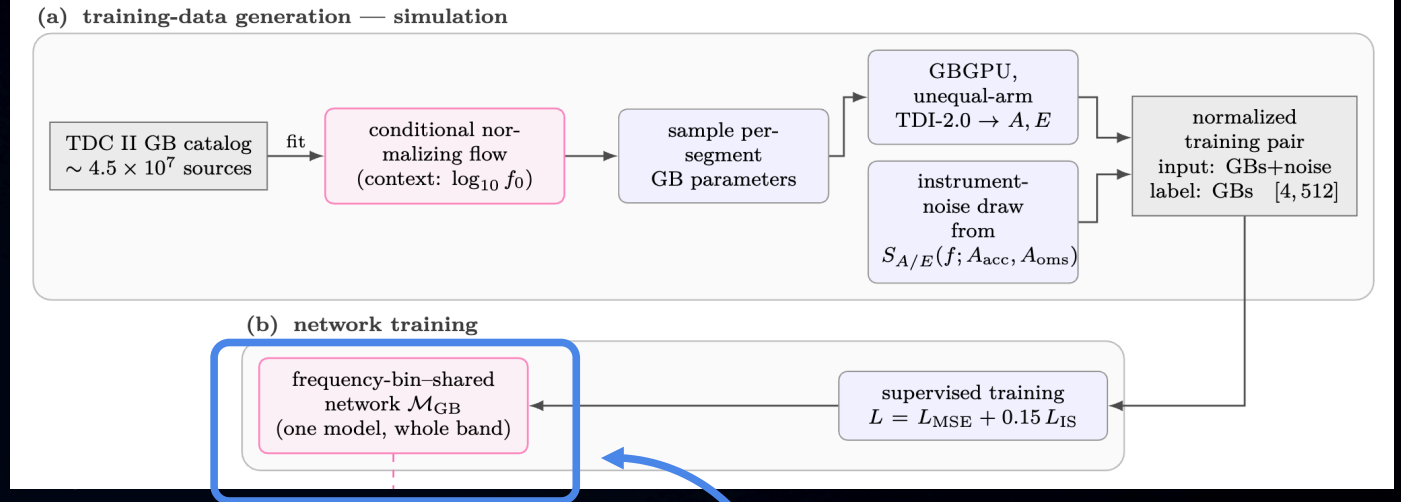
$$\mathcal{L}_{\text{IS}} = \frac{1}{M} \sum_{c,j} \left[\frac{R_{c,j}}{T_{c,j}} - \ln \frac{R_{c,j}}{T_{c,j}} - 1 \right]$$

where

- $R_{c,j} = \langle |d_c - \hat{g}_c|^2 \rangle_j$ is block-averaged residual power
- $T_{c,j} = \langle |n_c|^2 \rangle_j$ is block-averaged injected noise power
- j : eight coarse 64-bin sub-bands

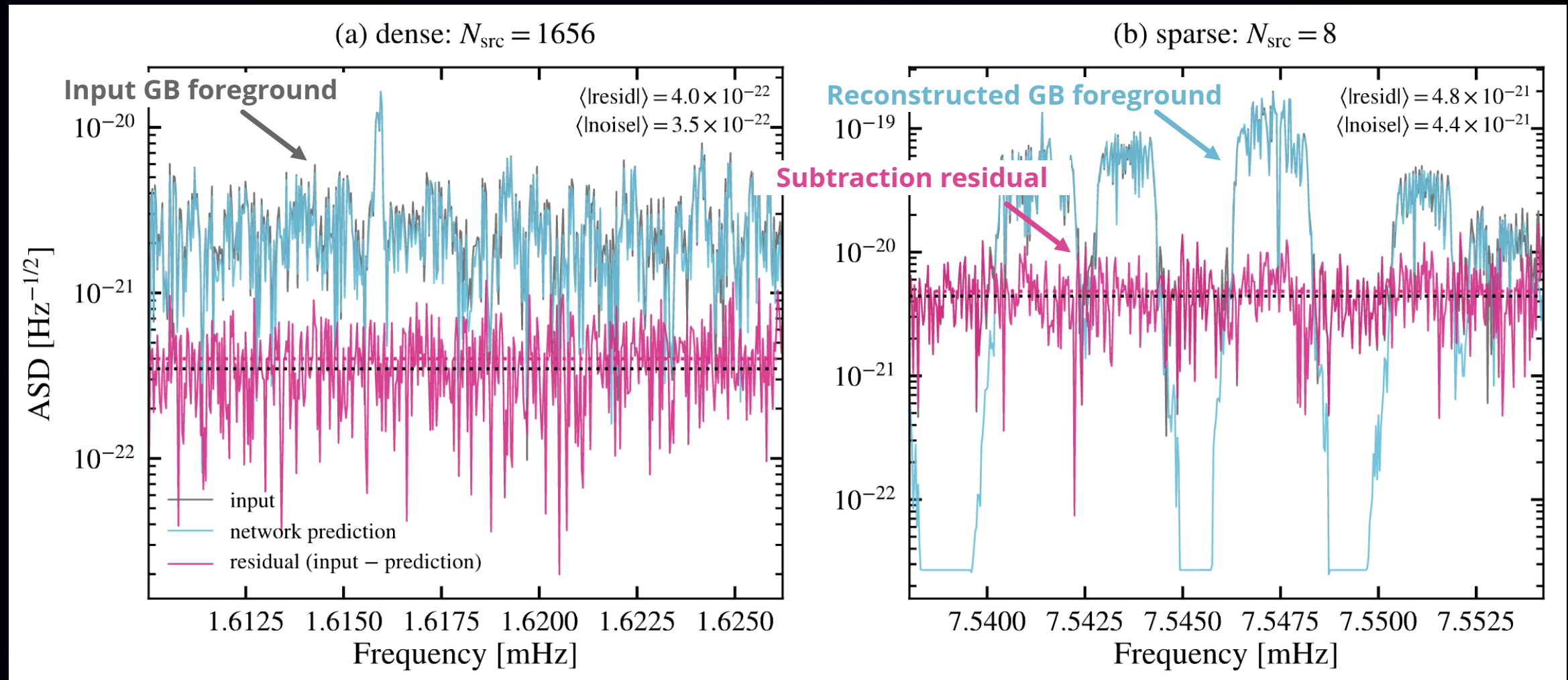


Loss landscape



One shared network generalizes across the full Galactic foreground

No retraining. No per-band tuning. One model everywhere.



The goal is not to identify every binary, but to recover a noise-dominated residual.

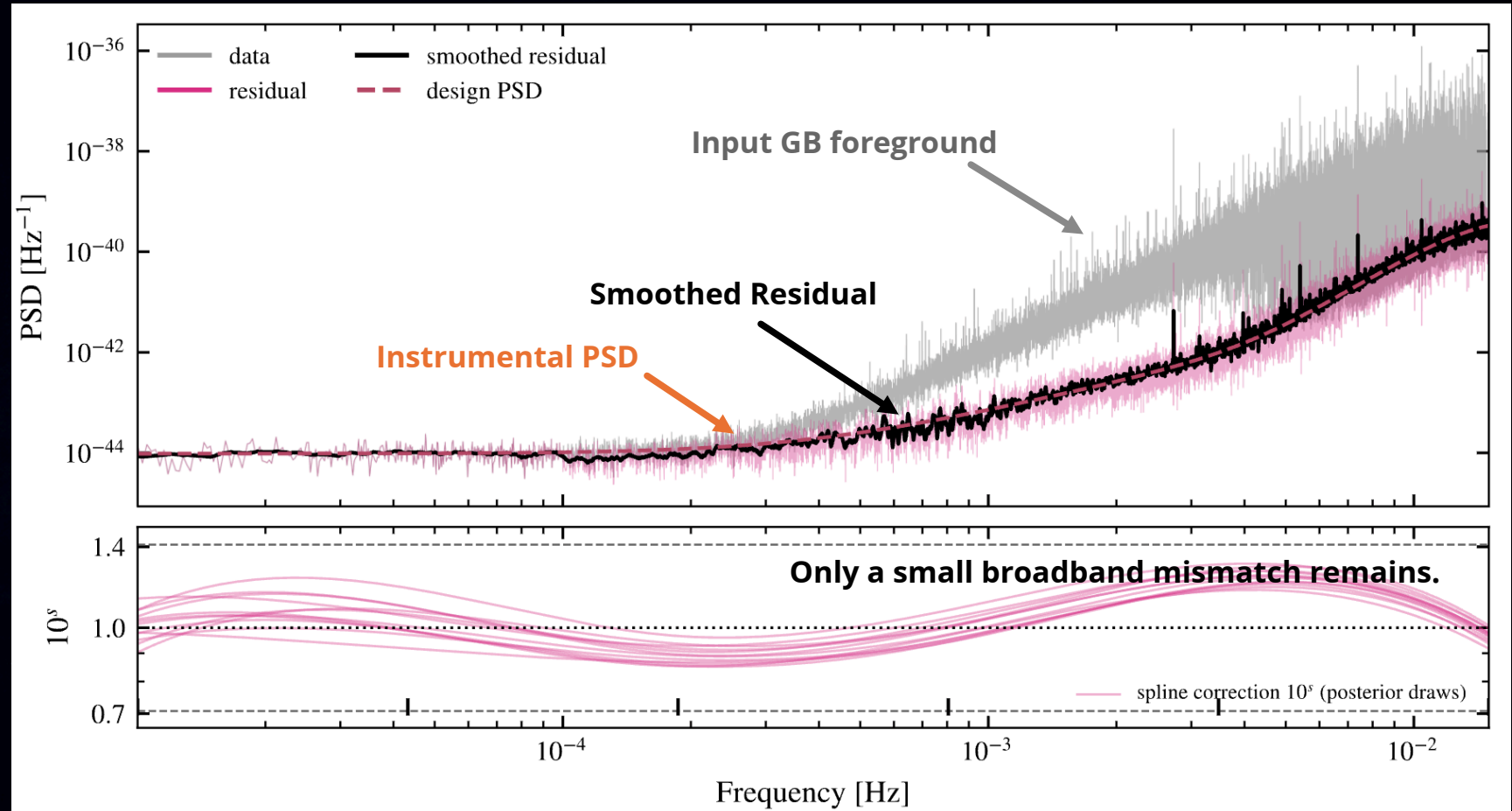
From 512-bin predictions to a continuous full-band residual

Sliding 512-bin windows are edge-cropped and stitched so that every frequency bin receives one reliable prediction.

Full-band inference

- 512-bin sliding windows
- 10% edge crop: 51 bins per side
- stride: 410 bins
- reflection padding at the band boundaries

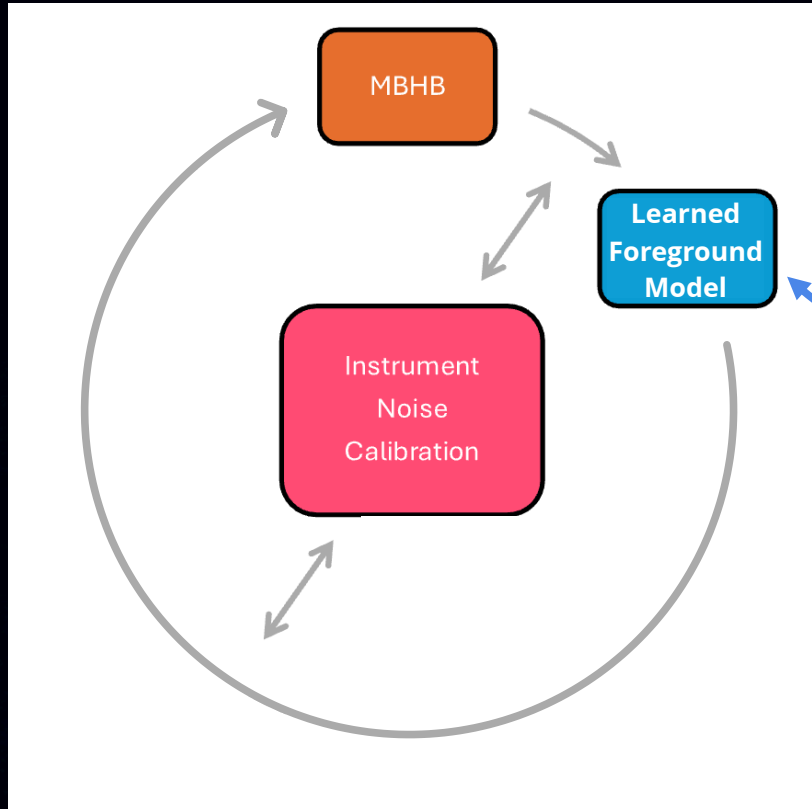
$$\hat{g}_c(f) = \text{stitch} \left[M_{\text{GB}}(d_c^{(w)}) \right]$$
$$r_c(f) = d_c(f) - \hat{g}_c(f)$$



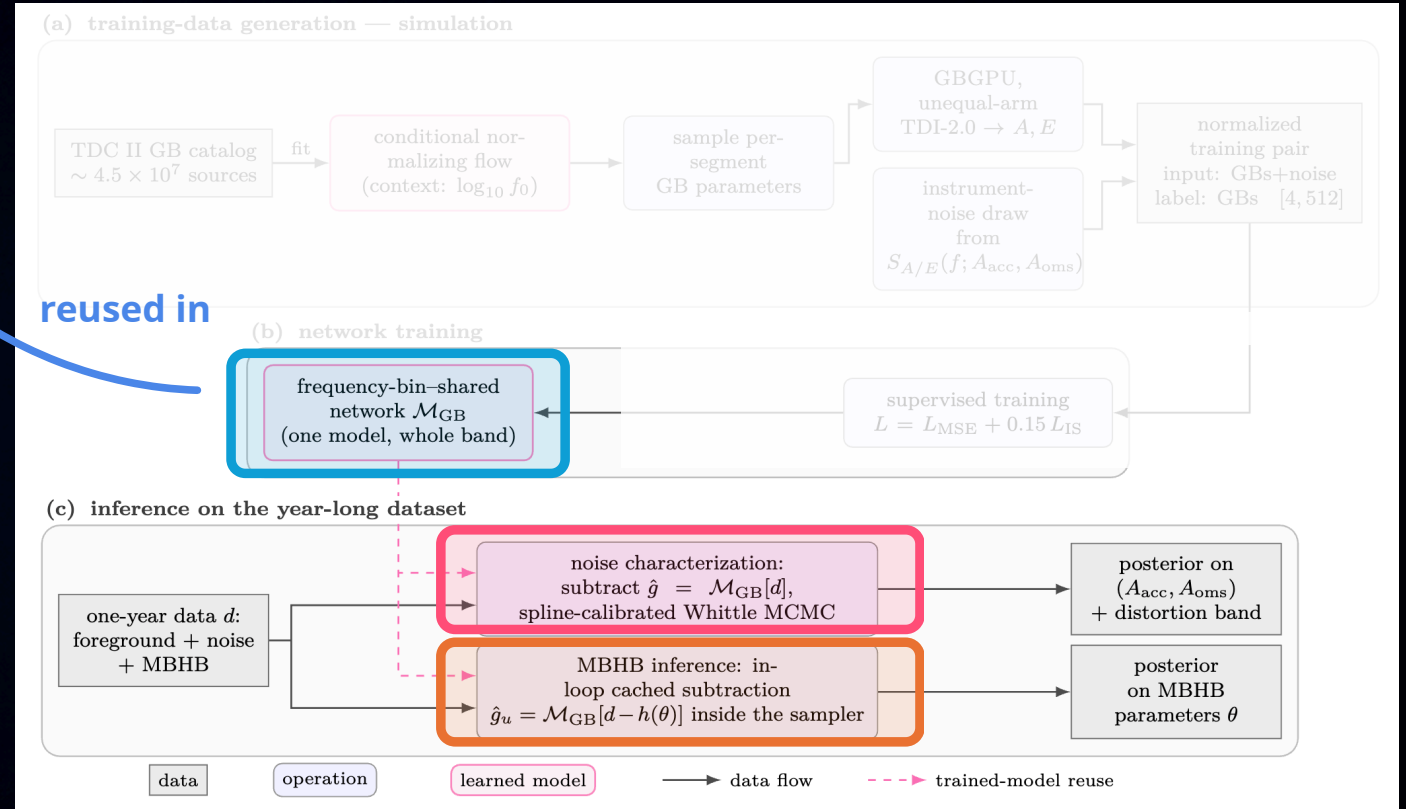
Local predictions become one continuous residual across the full analysis band.

The learned foreground enters Bayesian inference

The same foreground model is reused for both noise characterization and source inference.



Katz, Karnesis, Korsakova, Gair, and Stergioulas (2025)
(adapted)

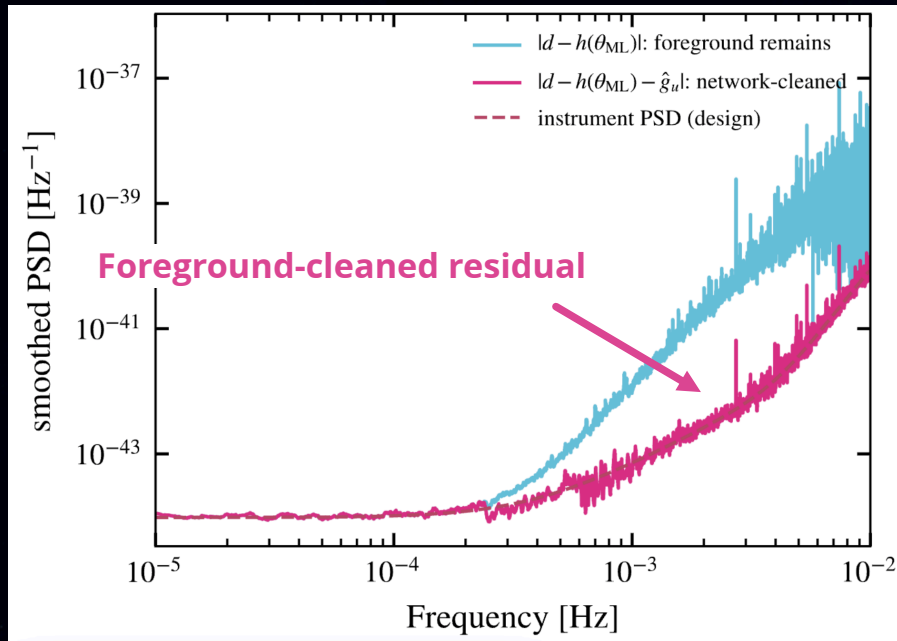


Pipeline overview

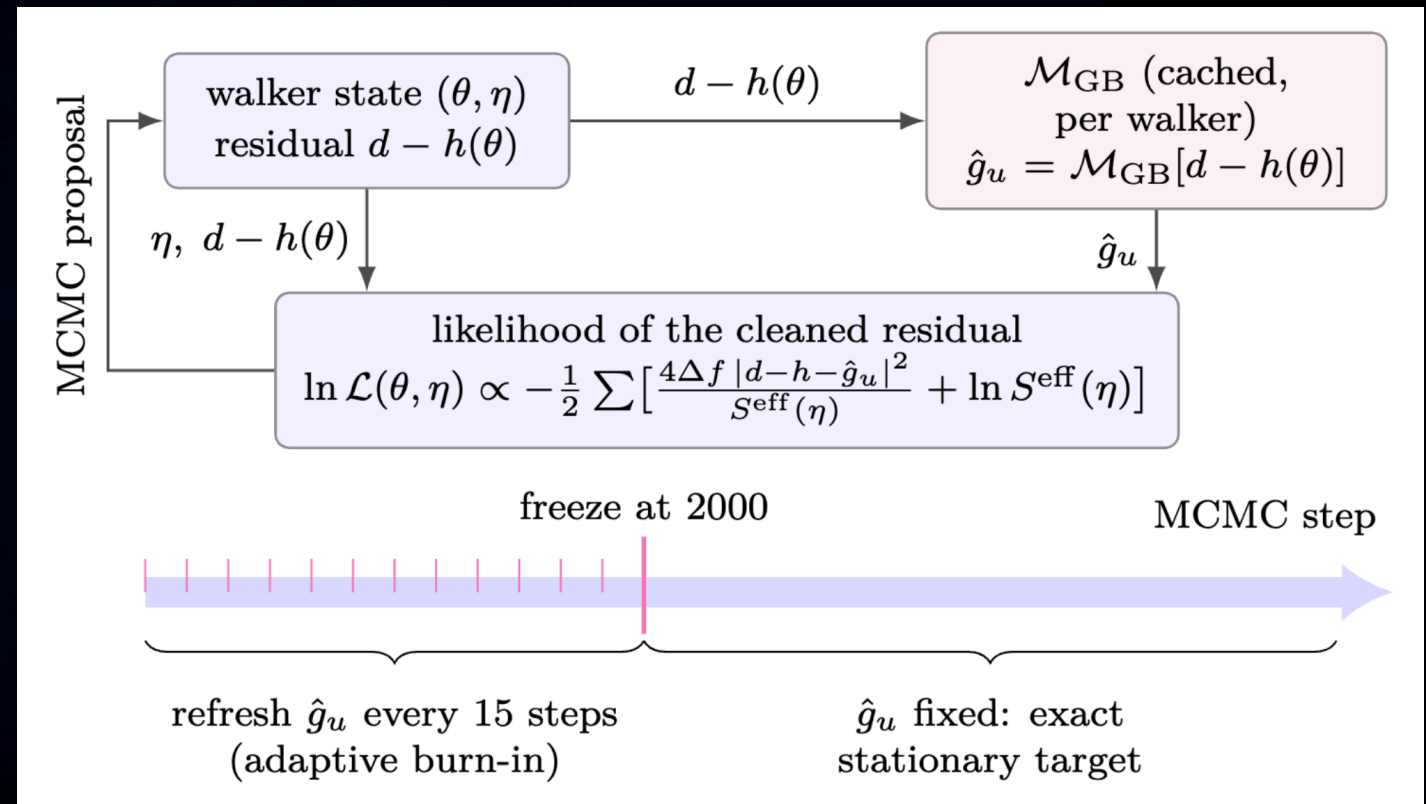
The learned foreground is reused—not retrained—inside downstream inference.

The learned foreground enters the likelihood

Every MCMC proposal is cleaned before evaluating the Whittle likelihood.



Every MCMC proposal is evaluated on the cleaned residual.



Model in the loop: The foreground prediction is refreshed every 15 iterations during burn-in and then frozen.

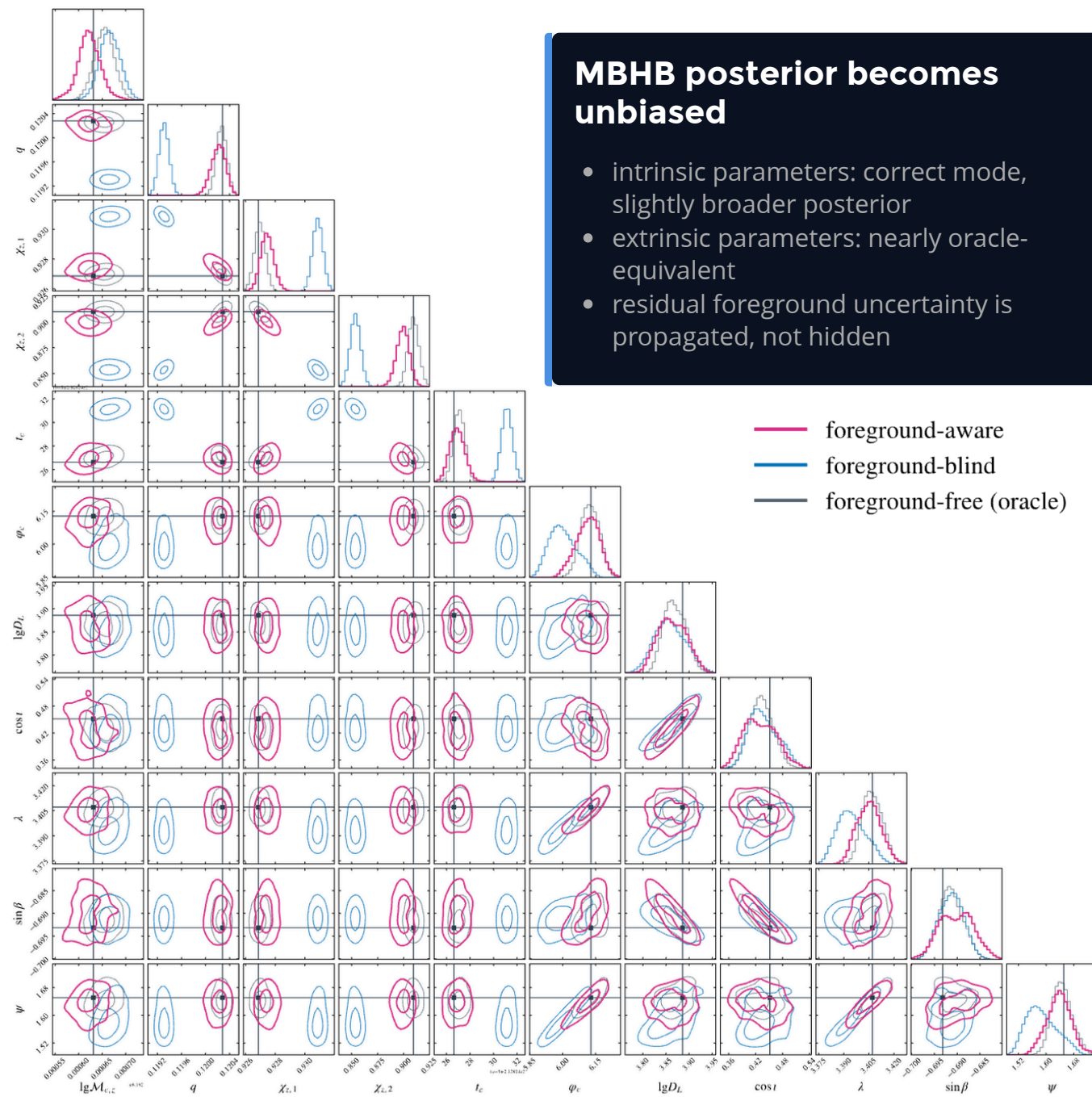
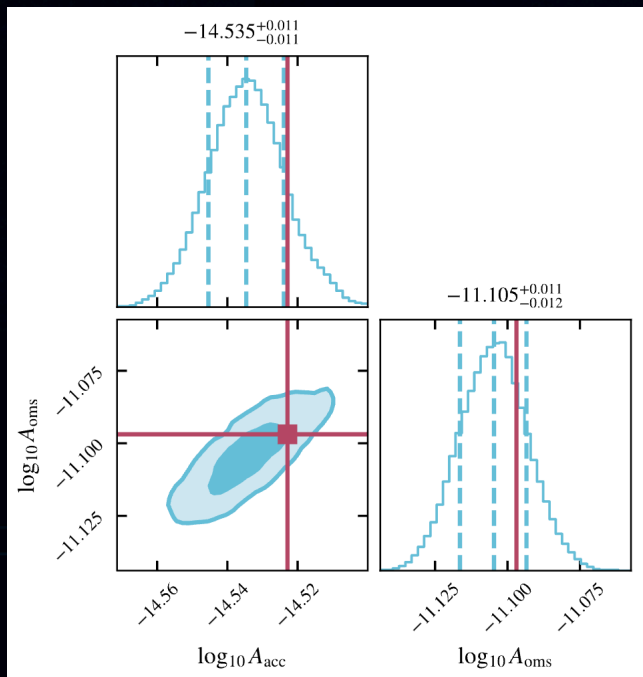
The sampler never fits the raw data — it fits the foreground-cleaned residual.

Foreground-aware inference removes bias from both the noise and the source posterior

Correcting the foreground first restores both instrumental-noise calibration and MBHB parameter estimation.

Noise characterization becomes unbiased

- recovered A_{acc} and A_{oms}
- posterior centered on the truth
- spline absorbs remaining smooth mismatch
- no systematic PSD inflation



MBHB posterior becomes unbiased

- intrinsic parameters: correct mode, slightly broader posterior
- extrinsic parameters: nearly oracle-equivalent
- residual foreground uncertainty is propagated, not hidden

Calibrating learned systematic uncertainty

The learned foreground is not assumed perfect—its remaining error is explicitly calibrated and propagated into the posterior.

Existing uncertainty

- detector noise
- finite observation
- finite measurement precision

New uncertainty from learning

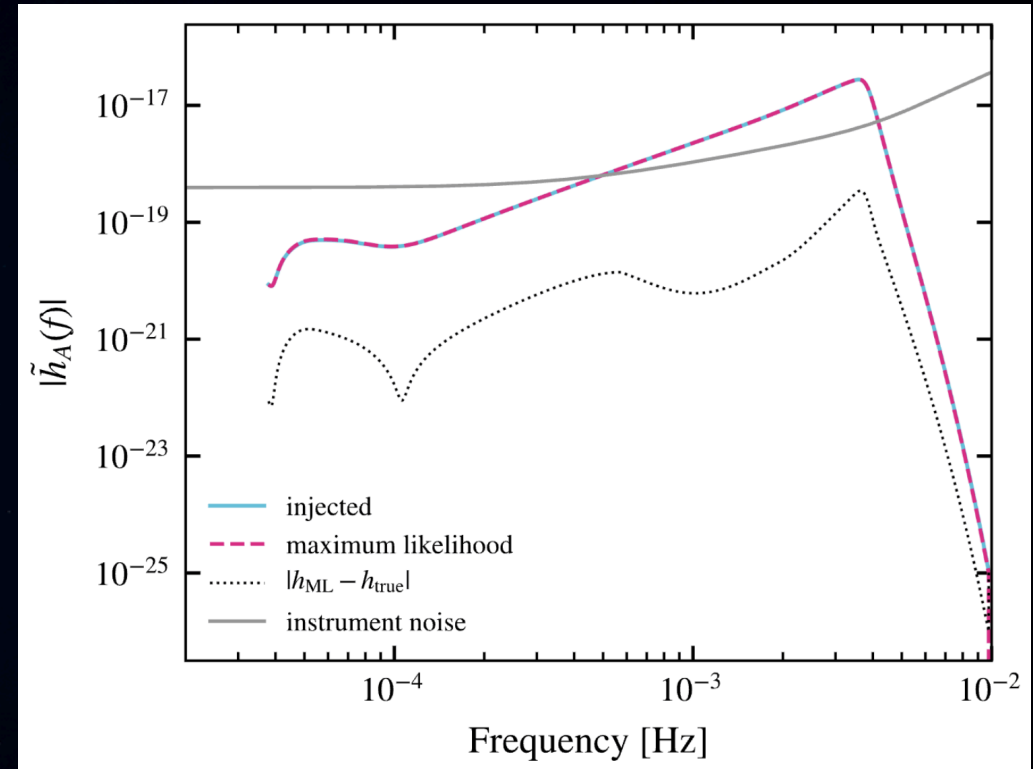
- imperfect foreground prediction
- residual foreground mismatch

↓

explicitly calibrated

Posterior interpretation

- bias largely removed
- remaining foreground uncertainty propagates into credible intervals
- intrinsic parameters become slightly broader
- no hidden over-confidence



The maximum-likelihood waveform matches the injection.
The learned foreground closely reproduces the aggregate Galactic signal, while leaving only instrument-like residuals.

Our goal is not to eliminate uncertainty, but to separate, calibrate, and propagate it honestly.

Bonus: toward EMRI-like targets

EMRIs are the natural next target because foreground bias accumulates over many cycles.

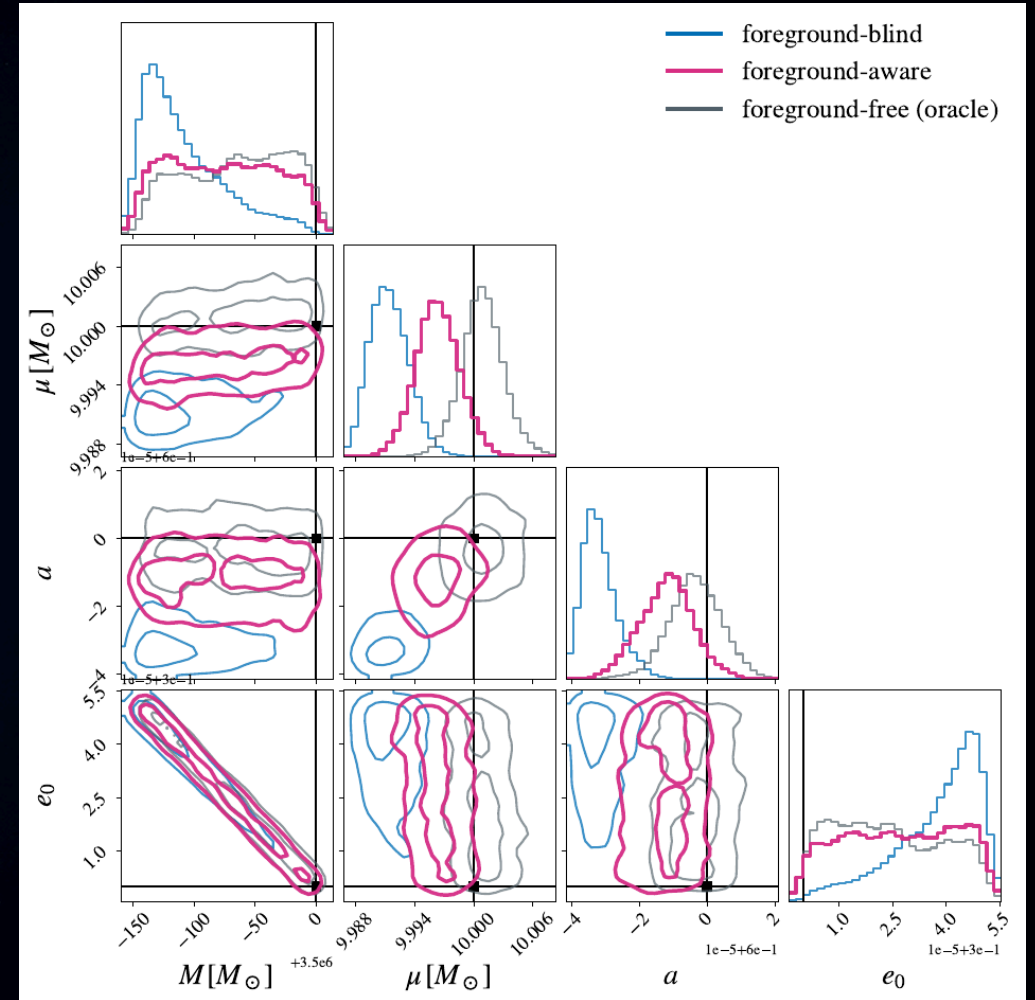
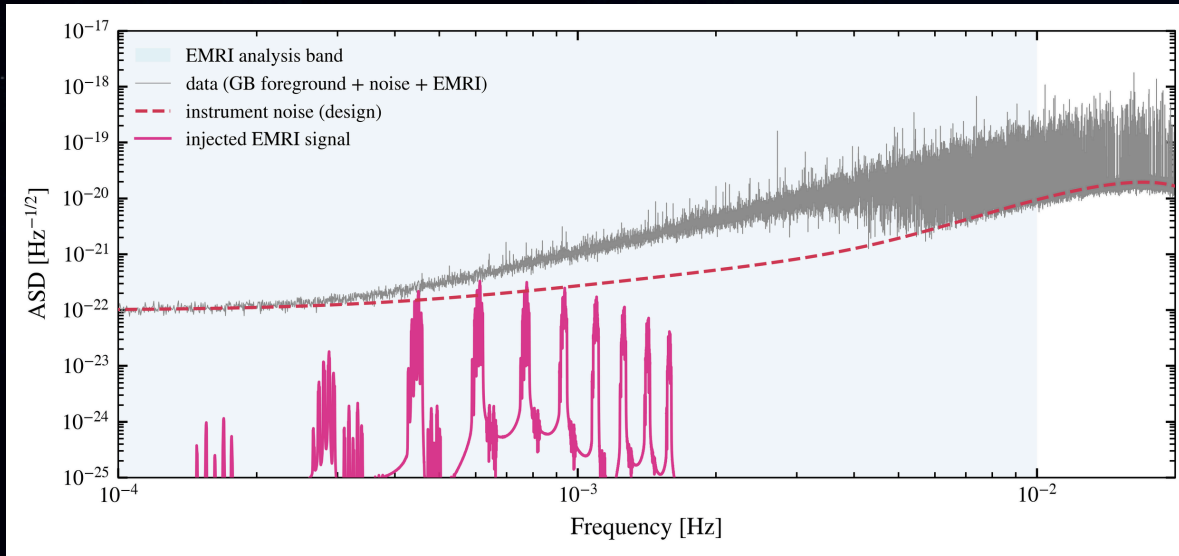
Toward weaker, longer-lived targets: the EMRI-like regime

EMRI-like target

- months to years
- 10^4 – 10^5 cycles
- low instantaneous SNR
- foreground bias accumulates coherently

Needed extensions:

- distributional $\hat{g}$
- pseudo-marginal / variance-inflated likelihood
- time-frequency conditioning
- phase-coherent validation



MBHB is the first hard test; EMRI is the real stress test.

Conclusions and outlook

Machine learning provides a practical and statistically honest bridge toward scalable LISA/Taiji global-fit inference.

Main contributions

- ✓ Learn the **aggregate Galactic foreground**, not individual binaries
- ✓ Recover **unbiased instrumental-noise calibration**
- ✓ Restore **MBHB posterior consistency** without source catalogs

Foreground estimation becomes an inference primitive.

What AI actually learned

AI learns NOT individual GB identities

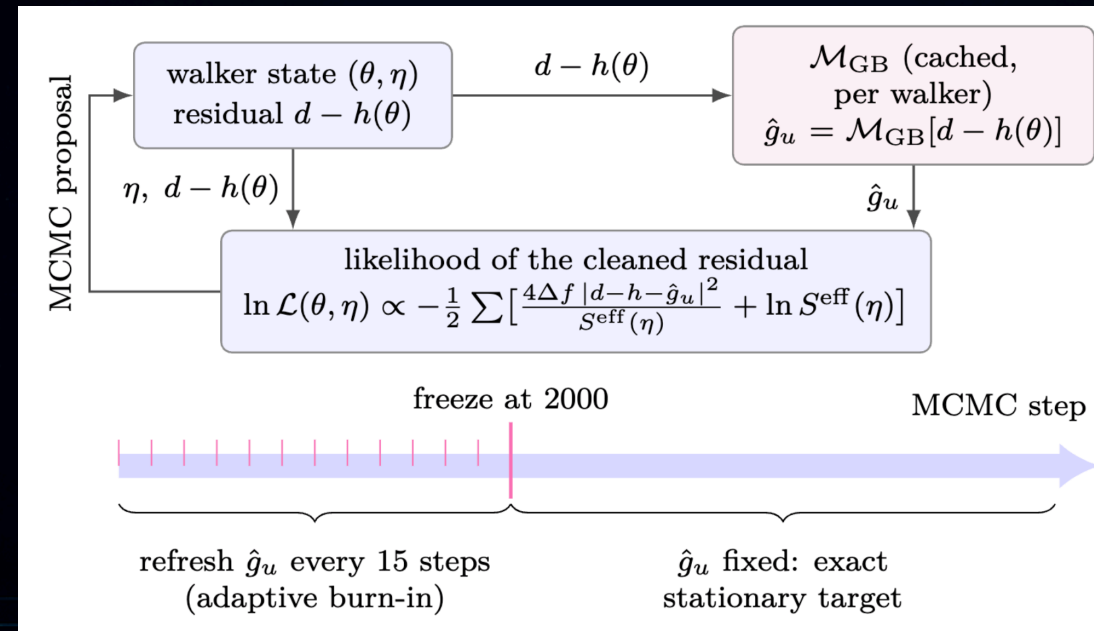


BUT a morphology → foreground map



usable residuals for downstream inference

Different question, complementary answer.



Not every Galactic binary must be identified before precision inference becomes possible.

Conclusions and outlook

Machine learning provides a practical and statistically honest bridge toward scalable LISA/Taiji global-fit inference.

```
for _ in range(num_of_audiences):  
    print('Thank you for your attention! 🙏')
```

Calibratable machine learning

- The systematic error of the learned component becomes a target of Bayesian calibration.

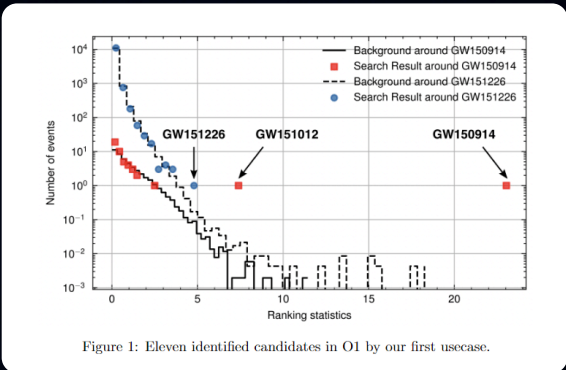
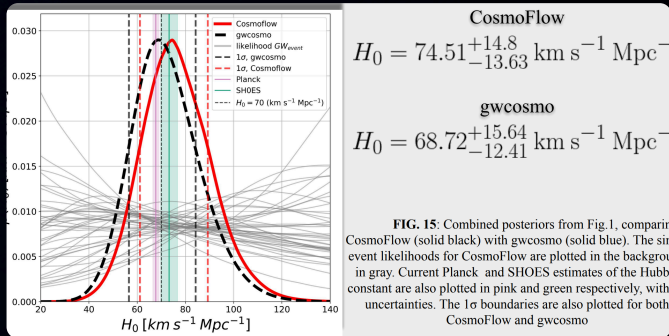


Figure 1: Eleven identified candidates in O1 by our first usecase.

HW+ (2024)



CosmoFlow
 $H_0 = 74.51^{+14.8}_{-13.63} \text{ km s}^{-1} \text{ Mpc}^{-1}$

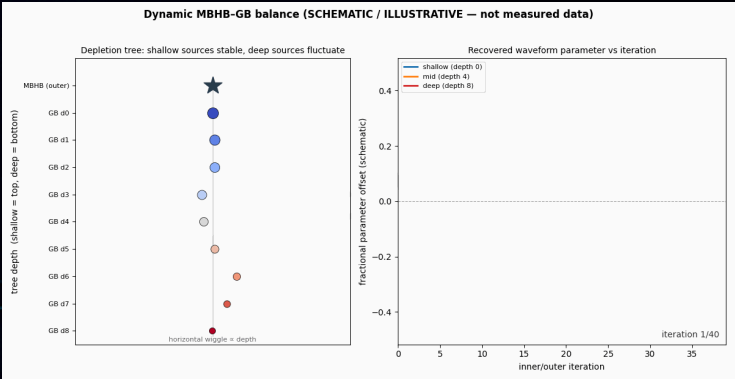
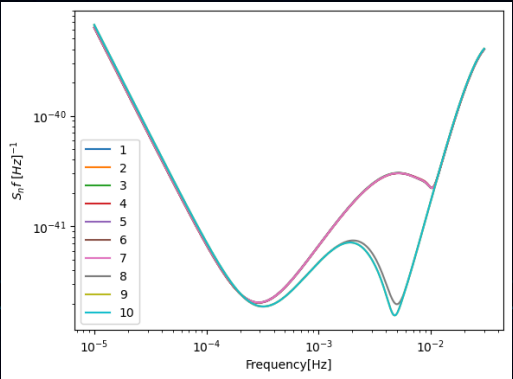
gwcosmo
 $H_0 = 68.72^{+15.64}_{-12.41} \text{ km s}^{-1} \text{ Mpc}^{-1}$

FIG. 15: Combined posteriors from Fig. 1, comparing CosmoFlow (solid black) with gwcosmo (solid blue). The search event likelihoods for CosmoFlow are plotted in the background in gray. Current Planck and SHOES estimates of the Hubble constant are also plotted in pink and green respectively, with uncertainties. The 1-sigma boundaries are also plotted for both CosmoFlow and gwcosmo.

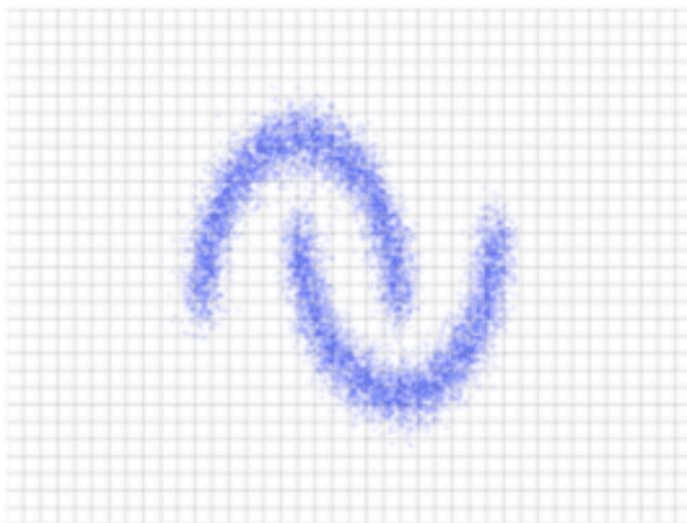
Stachurski+ (2024)

Ongoing work

1. Gee-Moo for TDCII (with Jianan Zhang, ICTP-AP)
2. MDP for global-fit (Frequentist vs Bayesian ...)
3. ...



Data space $\mathcal{X}$



Latent space $\mathcal{Z}$

