A Cross-Domain Simulator and Benchmark for Uncertainty Quantification in Global 21-cm Inference

Jacob Tutt[1] **Harry Bevins**[1,2] **James Alvey**[1] **Dominic Anstey**[1]
John Cumner[1] **Eloy de Lera Acedo**[1] **Anastasia Fialkov**[1]

[1]University of Cambridge, Cambridge CB3 0HA, UK
[2]Imperial College London, London SW7 2AZ, UK
jlt67@cam.ac.uk

Abstract

End-to-end uncertainty propagation in scientific inference often requires repeated evaluation of coupled simulators spanning multiple physical domains, often at prohibitive computational cost. Global 21-cm cosmology is an extreme example: a millikelvin cosmological signal must be separated from radio foregrounds and instrumental effects four orders of magnitude larger, making inference acutely sensitive to model misspecification. Reliable recovery is therefore contingent on an exceptionally accurate model that jointly propagates uncertainty through the complete observational chain. We present a differentiable, GPU-accelerated hybrid forward model combining neural emulation of the cosmological signal, a physics-informed multi-resolution parameterisation of the galactic foregrounds, and a linear surrogate model of the antenna. This unifies simulators spanning astrophysics and computational electromagnetics while reducing evaluation time by a factor of 7.2×10^7. Using GPU-accelerated Nested Slice Sampling, we benchmark the framework across 100 unseen signal realisations and find that 94 injected profiles lie entirely within their corresponding 95% credible bands. Together, the simulator and reference Nested Slice Sampling results provide a demanding benchmark of posterior calibration, mode coverage, and evidence accuracy across next-generation stochastic samplers, simulation-based inference, and variational inference frameworks.

1 Introduction

Observations of the first billion years of cosmic history remain sparse and indirect, leaving the astrophysical processes governing the formation of the first stars and galaxies weakly constrained by existing data. By tracing the thermal and ionisation state of the intergalactic medium (IGM), the redshifted 21-cm hyperfine transition of neutral hydrogen offers one of the most direct probes of this period [1, 2]. Single-antenna radiometer experiments [3–8] target its sky-averaged brightness temperature, $T_{21}(\nu)$, but realising their potential remains a formidable inference challenge [9–30].

Theoretical models predict the global signal to be of order a few hundred millikelvin [31], yet it must be extracted from Galactic and extragalactic foreground emission four orders of magnitude brighter [32]. Moreover, the antenna's direction- and frequency-dependent response to the radio sky, itself highly sensitive to geometric and environmental perturbations, non-trivially couples spatial foreground structure into the observed spectrum, producing features that are highly degenerate with the cosmological signal [33]. Crucially, neither the foreground nor instrumental contributions can be characterised *a priori* to the required precision, and modelling deficiencies in either can leave residuals far exceeding the signal itself [34–37]. Global 21-cm inference therefore operates in a severe

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Sim2Science: ML with Imperfect Scientific Models.

model-misspecification regime, in which an analysis may appear well constrained while recovering a systematically biased signal and a severely miscalibrated posterior.

Rigorous inference consequently requires a forward model that propagates uncertainty across the observational chain by coupling cross-domain simulators, from semi-numerical astrophysical simulations [38] to computational electromagnetic (CEM) antenna solvers [39]. The dimensionality of this uncertainty space compounds the cost of constructing large ensembles for simulation-based inference (SBI) [40] or repeatedly evaluating the model within Bayesian samplers. Forward-model acceleration addresses only one part of the computational challenge. The large dynamic range and strong degeneracies produce correlated, multimodal posterior geometries, posing a stringent test of sampling performance. These challenges motivate our accelerator-optimised hybrid framework. Implemented in JAX and coupled to GPU-native nested slice sampling (NSS) [41, 42], it enables end-to-end Bayesian inference over the global 21-cm observational chain at tractable computational cost.

Beyond its scientific application, this work addresses a timely methodological need: benchmarks must evolve with increasingly complex, high-dimensional scientific inverse problems. This 137-parameter problem provides a demanding open-source benchmark[1], with reference posterior samples and evidence estimates for evaluating next-generation Monte Carlo, simulation-based, and variational inference methods. The accelerated surrogate model is presented in Section 2 and statistically benchmarked in Section 3, with limitations and future work discussed in Section 4.

2 Hybrid Accelerated Forward Model

Figure 1 summarises the three primary uncertainty contributions required to construct physically informed global 21-cm observations. Even when generated concurrently, a full-fidelity CEM beam ($\sim$ 1 hour across 1,600 CPU cores) and astrophysical signal (2.8 hours across three cores) require $\sim$ 2.8 hours of wall time and 1,608 CPU-core hours (£16.08) per realisation.[2] Consequently, conventional analyses often simplify or fix poorly constrained components of the observational chain. We instead present a framework that accelerates each component independently, with their unified implementation and resulting computational acceleration quantified in Section 2.4.

2.1 Neural emulation of the global 21-cm signal

Simulating the global 21-cm signal requires the modelling of the radiative coupling, heating, and ionisation of the IGM and hence consideration of large-scale density and radiation fields alongside subgrid prescriptions for halo and star formation. We simulate these processes with the semi-numerical

[1]The code required to run the benchmark is publicly available on GitHub.

[2]Benchmarked using Intel Xeon Platinum 8368Q processors with 6.76 GB of RAM per core. Costs use the current CSD3 rate of £0.01 per CPU-core hour.

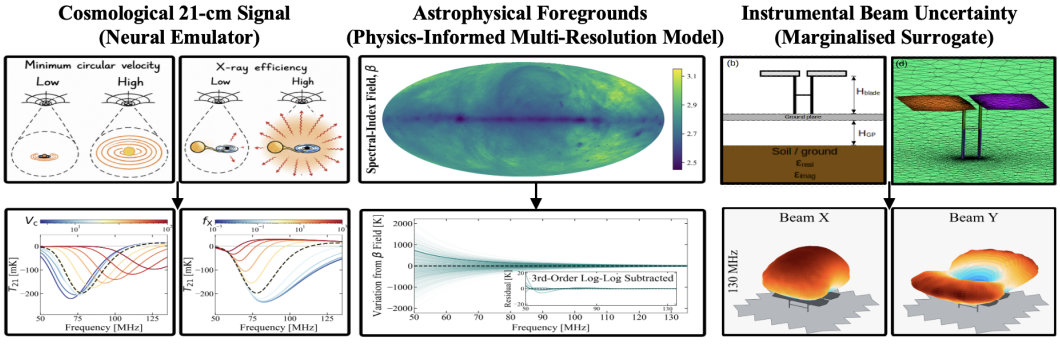


Figure 1: **Hybrid forward-model architecture.** Top: physical sources of uncertainty; bottom: their effect on the model. Left: a neural emulator maps astrophysical parameters to the global 21-cm signal. Centre: a physics-informed multi-resolution model represents uncertainty in the foreground sky. Right: a linear surrogate reconstructs chromatic beam variations from antenna and soil perturbations.

code `21cmSPACE` [31, 38], adopting the model of Reis et al. [43], parameterised by six astrophysical quantities, $\boldsymbol{\theta}_{21} = \{f_{\star,\mathrm{II}}, f_{\star,\mathrm{III}}, V_\mathrm{c}, f_\mathrm{X}, \tau, f_\mathrm{radio}\}$ (see Appendix A.1 for physical prescriptions).

To enable repeated evaluation during inference, we construct a neural emulator following Bevins et al. [44]. Rather than predicting a fixed discretised spectrum, we train it on `21cmSPACE` realisations with redshift supplied as a variable alongside the astrophysical parameters. This allows the nonlinear evolution of the global 21-cm signal to be represented by a compact network, while complete spectra across all redshifts can be evaluated in parallel. Across the test set, it achieves a RMSE of $0.91\ \mathrm{mK}$, more than an order of magnitude below the assumed $25\ \mathrm{mK}$ instrumental-noise level and sufficiently low to avoid appreciable bias in downstream inference [45]. Appendix A.2 specifies the dataset split, network architecture, training procedure, and evaluation underlying these benchmarks.

2.2 Physics-informed Foreground Model

Although the low-frequency radio sky is uncertain in both its absolute brightness and spectral structure, its broad spatial structure is well constrained by existing radio surveys and provides an informative physical template. We therefore parameterise regional departures from the Global Sky Model (GSM) reference map at $\nu_0 = 230\ \mathrm{MHz}$, $T_{230}(\Omega')$, in celestial coordinates Ω' [46], using independent sky partitions: N_α base-map correction regions with parameters α_i [47], and N_β spectral-index regions with parameters β_j [33], collectively $\boldsymbol{\theta}_\mathrm{FG} \equiv \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$. The regional construction and prior specification are provided in Appendix A.3. For given binary mask $M_{i,j}(\Omega')$ defining the intersection of the i-th base-map and j-th spectral-index region, the sky model is then:

$$T_\mathrm{sky}^\mathrm{model}(\Omega', \nu; \boldsymbol{\alpha}, \boldsymbol{\beta}) = [T_{230}(\Omega') - T_\mathrm{CMB}] \sum_{i=1}^{N_\alpha} \sum_{j=1}^{N_\beta} M_{i,j}(\Omega') \alpha_i \left(\frac{\nu}{\nu_0}\right)^{-\beta_j} + T_\mathrm{CMB}, \tag{1}$$

where T_CMB is the cosmic microwave background temperature.

2.3 Marginalised surrogate of beam chromaticity

High-fidelity CEM solvers [48–52] map a given antenna geometry and environmental conditions to a radiation pattern by solving the Electric Field Integral Equations. Since these physical parameters are purely instrumental nuisance quantities, we instead learn their induced beam uncertainty directly from 400 CEM realisations [53]. The surrogate must span the full beam-uncertainty space, encompassing variations of tens of kelvin, while retaining sensitivity to the millikelvin cosmological signal. We therefore introduce a hierarchical two-stage decomposition [53]. An initial singular value decomposition (SVD) extracts shared angular basis functions, $B_k(\Omega)$ [54], while a second decomposition captures correlations in their weights across frequency and angular mode, compressing the representation to just $M = 100$ coefficients. The resulting surrogate achieves a mean antenna-temperature reconstruction RMSE of $24.47\ \mathrm{mK}$, below the instrumental-noise level. The beam model can be written as

$$\tilde{D}(\nu, \Omega; \boldsymbol{a}) = \sum_{k=1}^{R} \left[e_{0,k}(\nu) + \sum_{m=1}^{M} a_m e_{m,k}(\nu) \right] B_k(\Omega), \tag{2}$$

where Ω denotes the antenna frame, R the number of retained angular modes, $e_{0,k}(\nu)$ the mean spectral weight, and $e_{m,k}(\nu)$ the correlated correction modes, weighted by the beam coefficients $\boldsymbol{\theta}_\mathrm{Beam} \equiv \boldsymbol{a} = \{a_m\}$. The CEM training set, surrogate construction, and validation are documented in Appendix A.4.

2.4 Accelerator-native implementation and performance

To maximise computational throughput, we precompute the time- and frequency-dependent coupling between foreground regions and beam modes for the fixed observing configuration as a reusable response operator:

$$\mathcal{B}_{i,j,m}(\nu, t) = \frac{1}{4\pi} \int_{4\pi} M_{i,j}(\Omega, t) \, [T_{230}(\Omega, t) - T_\mathrm{CMB}] \sum_{k=1}^{R} e_{m,k}(\nu) B_k(\Omega) \, \mathrm{d}\Omega, \tag{3}$$

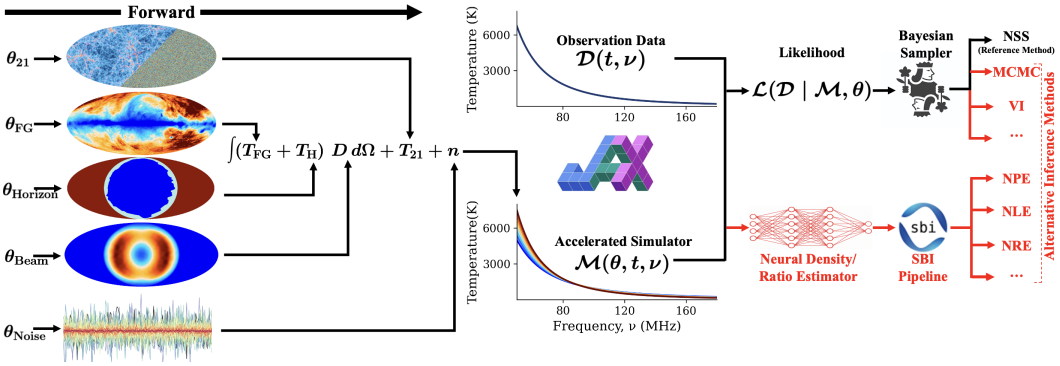


Figure 2: **Benchmark inference architecture.** Integration of full-fidelity observations and the accelerated simulator with the reference NSS analysis (black) and alternative inference methods (red).

where $M_{i,j}$ and T_{230} are rotated into the antenna frame at observing time t. The model antenna temperature, $T_{\rm m}$, then becomes (Figure 2)

$$T_{\rm m}(\nu, t) = \sum_{m=0}^{M} a_m \sum_{i=1}^{N_\alpha} \sum_{j=1}^{N_\beta} \alpha_i \mathcal{B}_{i,j,m}(\nu, t) \left(\frac{\nu}{\nu_0}\right)^{-\beta_j} + T_{\rm CMB} + T_{21}(\nu; \boldsymbol{\theta}_{21}) + n(\nu, t; \boldsymbol{\theta}_n), \quad (4)$$

where $a_0 \equiv 1$ selects the mean beam, T_{21} is the global 21-cm signal, and $n(\nu, t; \boldsymbol{\theta}_n)$ denotes a realisation of the adopted noise model, parameterised by $\boldsymbol{\theta}_{\rm Noise} \equiv \boldsymbol{\theta}_n$ [55].

End-to-end uncertainty propagation is thereby reduced to a compact neural-network evaluation and tensor contractions, yielding a low-memory, differentiable, and JIT-compilable forward model that is parallelisable across frequency, time, and simulations. On a single GPU, this implementation reduces the ~ 2.8-hour full-fidelity runtime to 0.140 ± 0.002 ms, a factor of 7.2×10^7.[3] Vectorising batches of 512 fully exploits GPU parallelism, reaching 384,939 simulations per second at a marginal cost of 2.1×10^{-9} pounds per simulation, a 7.6×10^9-fold reduction.

3 Benchmark Inference

Computational acceleration is only meaningful if the hybrid approximations adequately represent the relevant degeneracies, enabling accurate recovery and calibrated posterior uncertainty. At this dynamic range, residual surrogate errors can exceed the target signal, while inadequate physical constraints on nuisance components can leave it unidentifiable. We therefore benchmark end-to-end recovery using vectorised NSS [41, 42], jointly testing the fidelity of the accelerated simulator against full-fidelity observations and the sampler's exploration of the resulting posterior.[4] The benchmark comprises 100 unseen signal profiles spanning different levels of detectability and degeneracy with foreground and instrumental components. Each is injected into an observation generated using a full-fidelity sky realisation, a held-out CEM beam simulation, and a fixed three-hour observing window. The full formalism used to construct these datasets is detailed in Appendix A.5.

The complete forward model considered here contains 137 uncertain parameters: 100 beam coefficients, $\boldsymbol{\theta}_{\rm Beam}$, 30 foreground spectral indices, $\boldsymbol{\theta}_{\rm FG}$, six astrophysical parameters, $\boldsymbol{\theta}_{21}$, and one Gaussian noise amplitude, $\boldsymbol{\theta}_{\rm Noise}$. By design, the beam model is linear in $\boldsymbol{a}$ (Equation 2). Assuming the Gaussian prior $\boldsymbol{a} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_a)$, all 100 coefficients can thus be marginalised analytically, reducing the directly sampled space to 37 dimensions while retaining their contribution to the likelihood covariance:

$$\mathcal{L}_{\rm marg}(\mathbf{d} \mid \boldsymbol{\phi}) = \mathcal{N}\left[\mathbf{d} \,\middle|\, \boldsymbol{\mu}(\boldsymbol{\phi}), \mathbf{C}_n(\theta_n) + \mathbf{H}(\boldsymbol{\phi}) \boldsymbol{\Sigma}_a \mathbf{H}^{\rm T}(\boldsymbol{\phi})\right], \quad (5)$$

where $\mathbf{d}$ is the observed data vector, $\mathbf{C}_n$ the noise covariance, $\boldsymbol{\phi} = (\boldsymbol{\theta}_{\rm FG}, \boldsymbol{\theta}_{21}, \boldsymbol{\theta}_{\rm Noise})$, $\boldsymbol{\mu}$ is the mean-beam forward model, and $\mathbf{H}$ maps the beam coefficients into the observed spectrum (derived in Appendix A.6). Using this covariance form, the complete marginalised likelihood takes 0.291 ± 0.003 ms per evaluation and reaches 30,696 evaluations per second in batches of 2,048.[3]

[3] Timings report the mean $\pm$ standard deviation across 500 warm evaluations on a single NVIDIA H100 GPU with 96 GB of HBM3. Costs assume £2.94 per H100-hour based on AWS pricing.

[4] The synthetic benchmark datasets, NSS chains, and inference products are available in Zenodo record 22046858.

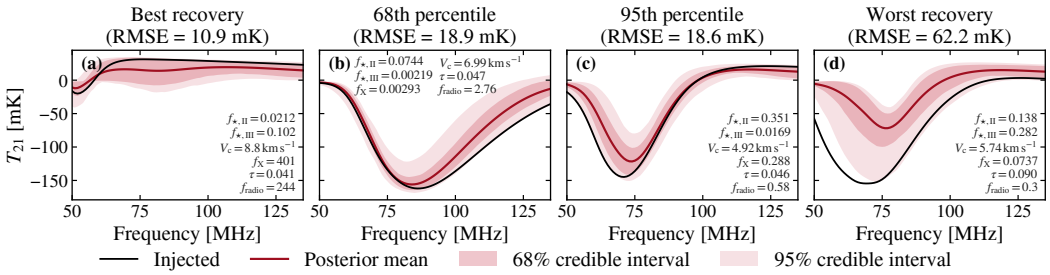


Figure 3: **Signal-profile recovery.** Representative cases ordered by minimum whole-profile credible level. Lines denote the injection and posterior mean, shading the pointwise 68% and 95% intervals, and titles the RMSE.

3.1 Signal recovery

We evaluate recovery in signal space, since the intrinsic degeneracies and multimodality of the astrophysical parameterisation broaden the marginal parameter posteriors (Appendix A.9) and would otherwise be conflated with inference performance. Signal-profile credible-band coverage evaluates posterior calibration, while posterior-mean RMSE independently quantifies the accuracy of the recovered signal, with both diagnostics defined in Appendix A.8.

Figure 3 shows representative profile recoveries. Across the complete benchmark, 94 of 100 injected profiles lie entirely within their 95% whole-profile credible bands, consistent with nominal coverage, and 75 posterior-mean profiles have an RMSE below the assumed 25 mK instrumental-noise level. Residual departures between the posterior mean and injected profile are therefore consistent with calibrated uncertainty, with the credible-band widths reflecting genuine degeneracies between the signal and the foreground and beam components. Taken together, these results show that the computational gains of the hybrid framework are achieved without compromising signal recovery or posterior calibration, even when no component of the observational chain is treated as known. The open-source benchmark therefore provides full-fidelity test observations, the precomputed response operator, and ready-to-run implementations of the accelerated simulator and its full and analytically marginalised likelihoods, forming a 100-realisation benchmark suite for alternative inference methods. Reference NSS posteriors and evidence estimates are released for all 100 realisations.

4 Limitations and Future Work

To our knowledge, this is the first work to jointly propagate the dominant uncertainties in global 21-cm inference in a physically motivated Bayesian forward-modelling framework, but a complete digital twin must also incorporate secondary observational effects [56–62], receiver-calibration uncertainty [63, 64], and radiometer noise beyond the homoskedastic Gaussian model adopted here. Existing formalisms support these extensions: additional physical components can be integrated modularly, while heteroskedastic noise is accommodated through a diagonal $\mathbf{C}_n$ set by the radiometer equation. These will nevertheless increase dimensionality and complicate the posterior geometry. This exacerbates the remaining inference bottleneck: NSS scales theoretically as $\mathcal{O}(D^3)$ [65], while the present 37-dimensional analyses already require 5 h 34 min to 7 h 16 min per realisation,[5] making scalable posterior exploration and evidence-based model comparison a key challenge [28].

While standard benchmark suites such as `sbibm` support controlled and reproducible comparison of inference methods [66], scientific applications are increasing in dimensionality and structural complexity. Examples include the transition from compact gravitational-wave source models to stochastic backgrounds in astrophysics [67], and from point-neuron models to high-dimensional inference over detailed biophysical neurons and circuits in neuroscience [68–70]. Designed for direct integration and rapid, low-cost evaluation, the framework supports benchmarking next-generation sampling algorithms [71, 72], simulation-based inference [40, 73–77], and variational inference [78, 79], testing computational efficiency, posterior calibration, mode coverage, and evidence accuracy on a scientifically timely, high-dimensional inverse problem at the frontier of early-Universe cosmology.

[5] Totalling 594.4 GPU-hours across the 100 runs, each on one NVIDIA H100 GPU (96 GB HBM3).

Acknowledgments and Disclosure of Funding

The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI [80] is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023]. This work used the DiRAC Data Intensive service (CSD3, project number APP52749) at the University of Cambridge, managed by the University of Cambridge University Information Services on behalf of the STFC DiRAC HPC Facility (`www.dirac.ac.uk`). The DiRAC component of CSD3 at Cambridge was funded by BEIS, UKRI and STFC capital funding and STFC operations grants. DiRAC is part of the UKRI Digital Research Infrastructure. JT is supported by the Harding Distinguished Postgraduate Scholars Programme (HDPSP) and the Science and Technology Facilities Council (STFC) DTP Studentship. HTJB and JA acknowledge support from the Kavli Institute for Cosmology Cambridge and the Kavli Foundation. HTJB also acknowledges support from Imperial College London and Schmidt Sciences through an Imperial College Research Fellowship. JMC and DJA were supported by the Science and Technology Facilities Council. We would also like to thank the Kavli Foundation for their support of REACH. EdLA acknowledges the support of STFC via an Ernest Rutherford Fellowship. This work was supported by UK Research and Innovation (UKRI) through the Horizon Europe Guarantee under grant EP/Y02916X/1 (REACH 21).

References

[1] Steven R. Furlanetto, S. Peng Oh, and Frank H. Briggs. Cosmology at low frequencies: The 21cm transition and the high-redshift universe. *Physics Reports*, 433(4–6):181–301, October 2006. ISSN 0370-1573. doi: 10.1016/j.physrep.2006.08.002. URL `http://dx.doi.org/10.1016/j.physrep.2006.08.002`.

[2] Jonathan R Pritchard and Abraham Loeb. 21 cm cosmology in the 21st century. *Reports on Progress in Physics*, 75(8):086901, July 2012. ISSN 1361-6633. doi: 10.1088/0034-4885/75/8/086901. URL `http://dx.doi.org/10.1088/0034-4885/75/8/086901`.

[3] Judd D. Bowman, Alan E. E. Rogers, Raul A. Monsalve, Thomas J. Mozdzen, and Nivedita Mahesh. An absorption profile centred at 78 megahertz in the sky-averaged spectrum. *Nature*, 555(7694):67–70, March 2018. ISSN 1476-4687. doi: 10.1038/nature25792. URL `http://dx.doi.org/10.1038/nature25792`.

[4] Liju Philip, Z. Abdurashidova, H. Chiang, N. Ghazi, A. Gumba, H. Heilgendorff, Jose Jáuregui Garcia, Kagiso Malepe, C. Nunhokee, Jaron Peterson, J. Sievers, V. Simes, and R. Spann. Probing radio intensity at High-Z from Marion: 2017 instrument. *Journal of Astronomical Instrumentation*, 08(2):1950004, 2019. doi: 10.1142/S2251171719500041.

[5] Saurabh Singh, Ravi Subrahmanyan, N. Udaya Shankar, Mayuri Sathyanarayana Rao, Anastasia Fialkov, Aviad Cohen, Rennan Barkana, B. S. Girish, A. Raghunathan, R. Somashekar, and K. S. Srivani. SARAS 2 Constraints on Global 21 cm Signals from the Epoch of Reionization. *The Astrophysical Journal*, 858(1):54, May 2018. doi: 10.3847/1538-4357/aabae1.

[6] Raul A. Monsalve, Christian H. Bye, Jonathan L. Sievers, Vadym Bidula, Ricardo Bustos, H. Cynthia Chiang, Xinze Guo, Ian Hendricksen, Francis McGee, F. Patricio Mena, Garima Prabhakar, Oscar Restrepo, and Nithyanandan Thyagarajan. Simulating the detection of the global 21 cm signal with MIST for different models of the soil and beam directivity. *The Astrophysical Journal*, 961(1):56, January 2024. ISSN 1538-4357. doi: 10.3847/1538-4357/ad0f1b. URL `http://dx.doi.org/10.3847/1538-4357/ad0f1b`.

[7] Philip Bull, Ahmed El-Makadema, Hugh Garsden, John Edgley, Neil Roddis, Jens Chluba, Christopher J Conselice, Sohini Dutta, Katrine A Glasscock, Ainulnabilah Nasirudin, Jordan Norris, Michael J Wilensky, Isabelle Ye, and Zheng Zhang. RHINO: a large horn antenna for detecting the 21 cm global signal. *RAS Techniques and Instruments*, 4:rzaf046, 2025. ISSN 2752-8200. doi: 10.1093/rasti/rzaf046. URL `http://dx.doi.org/10.1093/rasti/rzaf046`.

[8] E. de Lera Acedo, D. I. L. de Villiers, N. Razavi-Ghods, W. Handley, A. Fialkov, A. Magro, D. Anstey, H. T. J. Bevins, R. Chiello, J. Cumner, A. T. Josaitis, I. L. V. Roque, P. H. Sims, K. H. Scheutwinkel, P. Alexander, G. Bernardi, S. Carey, J. Cavillot, W. Croukamp, J. A. Ely, T. Gessey-Jones, Q. Gueuning, R. Hills, G. Kulkarni, R. Maiolino, P. D. Meerburg, S. Mittal, J. R.

Pritchard, E. Puchwein, A. Saxena, E. Shen, O. Smirnov, M. Spinelli, and K. Zarb-Adami. The REACH radiometer for detecting the 21-cm hydrogen signal from redshift z ≈ 7.5–28. *Nature Astronomy*, 6(8):984–998, July 2022. ISSN 2397-3366. doi: 10.1038/s41550-022-01709-9. URL http://dx.doi.org/10.1038/s41550-022-01709-9.

[9] Geraint J. A. Harker, Jonathan R. Pritchard, Jack O. Burns, and Judd D. Bowman. An MCMC approach to extracting the global 21-cm signal during the cosmic dawn from sky-averaged radio observations. *Monthly Notices of the Royal Astronomical Society*, 419(2):1070–1084, January 2012. doi: 10.1111/j.1365-2966.2011.19766.x.

[10] Adrian Liu, Jonathan R. Pritchard, Max Tegmark, and Abraham Loeb. Global 21 cm signal experiments: A designer's guide. *Physical Review D*, 87(4):043002, February 2013. ISSN 1550-2368. doi: 10.1103/physrevd.87.043002. URL http://dx.doi.org/10.1103/PhysRevD.87.043002.

[11] Eric R. Switzer and Adrian Liu. Erasing the variable: Empirical foreground discovery for global 21 cm spectrum experiments. *The Astrophysical Journal*, 793(2):102, September 2014. ISSN 1538-4357. doi: 10.1088/0004-637x/793/2/102. URL http://dx.doi.org/10.1088/0004-637X/793/2/102.

[12] Raul A. Monsalve, Alan E. E. Rogers, Judd D. Bowman, and Thomas J. Mozdzen. Results from edges high-band. i. constraints on phenomenological models for the global 21 cm signal. *The Astrophysical Journal*, 847(1):64, September 2017. doi: 10.3847/1538-4357/aa88d1. URL https://doi.org/10.3847/1538-4357/aa88d1.

[13] Keith Tauscher, David Rapetti, Jack O. Burns, and Eric Switzer. Global 21 cm signal extraction from foreground and instrumental effects. I. pattern recognition framework for separation using training sets. *The Astrophysical Journal*, 853(2):187, February 2018. ISSN 1538-4357. doi: 10.3847/1538-4357/aaa41f. URL http://dx.doi.org/10.3847/1538-4357/aaa41f.

[14] K. Tauscher, D. Rapetti, and J.O. Burns. A new goodness-of-fit statistic and its application to 21-cm cosmology. *Journal of Cosmology and Astroparticle Physics*, 2018(12):015–015, December 2018. ISSN 1475-7516. doi: 10.1088/1475-7516/2018/12/015. URL http://dx.doi.org/10.1088/1475-7516/2018/12/015.

[15] Richard Hills, Girish Kulkarni, P. Daniel Meerburg, and Ewald Puchwein. Concerns about modelling of the edges data. *Nature*, 564(7736):E32–E34, December 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0796-5. URL http://dx.doi.org/10.1038/s41586-018-0796-5.

[16] Saurabh Singh and Ravi Subrahmanyan. The redshifted 21 cm signal in the EDGES low-band spectrum. *The Astrophysical Journal*, 880(1):26, July 2019. ISSN 1538-4357. doi: 10.3847/1538-4357/ab2879. URL http://dx.doi.org/10.3847/1538-4357/ab2879.

[17] Peter H. Sims and Jonathan C. Pober. Testing for calibration systematics in the EDGES low-band data using Bayesian model selection. *Monthly Notices of the Royal Astronomical Society*, 492(1):22–38, February 2020. doi: 10.1093/mnras/stz3388.

[18] David Rapetti, Keith Tauscher, Jordan Mirocha, and Jack O. Burns. Global 21 cm signal extraction from foreground and instrumental effects. II. efficient and self-consistent technique for constraining nonlinear signal models. *The Astrophysical Journal*, 897(2):174, July 2020. ISSN 1538-4357. doi: 10.3847/1538-4357/ab9b29. URL http://dx.doi.org/10.3847/1538-4357/ab9b29.

[19] Keith Tauscher, David Rapetti, and Jack O. Burns. Global 21 cm signal extraction from foreground and instrumental effects. III. utilizing drift-scan time dependence and full stokes measurements. *The Astrophysical Journal*, 897(2):175, July 2020. ISSN 1538-4357. doi: 10.3847/1538-4357/ab9b2a. URL http://dx.doi.org/10.3847/1538-4357/ab9b2a.

[20] Keith Tauscher, David Rapetti, and Jack O. Burns. Formulating and critically examining the assumptions of global 21 cm signal analyses: How to avoid the false troughs that can appear in single-spectrum fits. *The Astrophysical Journal*, 897(2):132, July 2020. ISSN 1538-4357. doi: 10.3847/1538-4357/ab9a3f. URL http://dx.doi.org/10.3847/1538-4357/ab9a3f.

[21] Keith Tauscher, David Rapetti, Bang D. Nhan, Alec Handy, Neil Bassett, Joshua Hibbard, David Bordenave, Richard F. Bradley, and Jack O. Burns. Global 21 cm signal extraction from foreground and instrumental effects. IV. accounting for realistic instrument uncertainties and their overlap with foreground and signal models. *The Astrophysical Journal*, 915(1):66, July

2021. ISSN 1538-4357. doi: 10.3847/1538-4357/ac00af. URL http://dx.doi.org/10.3847/1538-4357/ac00af.

[22] H T J Bevins, W J Handley, A Fialkov, E de Lera Acedo, L J Greenhill, and D C Price. <scp>maxsmooth</scp>: rapid maximally smooth function fitting with applications in global 21-cm cosmology. *Monthly Notices of the Royal Astronomical Society*, 502(3):4405–4425, January 2021. ISSN 1365-2966. doi: 10.1093/mnras/stab152. URL http://dx.doi.org/10.1093/mnras/stab152.

[23] Dominic Anstey, Eloy de Lera Acedo, and Will Handley. Use of time dependent data in bayesian global 21-cm foreground and signal modelling. *Monthly Notices of the Royal Astronomical Society*, 520(1):850–865, 01 2023. ISSN 0035-8711. doi: 10.1093/mnras/stad156. URL https://doi.org/10.1093/mnras/stad156.

[24] Steven G. Murray, Judd D. Bowman, Peter H. Sims, Nivedita Mahesh, Alan E. E. Rogers, Raul A. Monsalve, Titu Samson, and Akshatha Konakondula Vydula. A bayesian calibration framework for EDGES. *Monthly Notices of the Royal Astronomical Society*, 517(2):2264–2284, September 2022. ISSN 1365-2966. doi: 10.1093/mnras/stac2600. URL http://dx.doi.org/10.1093/mnras/stac2600.

[25] Anchal Saxena, P Daniel Meerburg, Eloy de Lera Acedo, Will Handley, and Léon V E Koopmans. Sky-averaged 21-cm signal extraction using multiple antennas with an SVD framework: the REACH case. *Monthly Notices of the Royal Astronomical Society*, 522 (1):1022–1032, April 2023. ISSN 1365-2966. doi: 10.1093/mnras/stad1047. URL http://dx.doi.org/10.1093/mnras/stad1047.

[26] Joshua J. Hibbard, David Rapetti, Jack O. Burns, Nivedita Mahesh, and Neil Bassett. Fitting and comparing galactic foreground models for unbiased 21 cm cosmology. *The Astrophysical Journal*, 959(2):103, December 2023. ISSN 1538-4357. doi: 10.3847/1538-4357/ad06b3. URL http://dx.doi.org/10.3847/1538-4357/ad06b3.

[27] Anchal Saxena, P Daniel Meerburg, Christoph Weniger, Eloy de Lera Acedo, and Will Handley. Simulation-based inference of the sky-averaged 21-cm signal from CD-EoR with REACH. *RAS Techniques and Instruments*, 3(1):724–736, Jan 2024. ISSN 2752-8200. doi: 10.1093/rasti/rzae047. URL http://dx.doi.org/10.1093/rasti/rzae047.

[28] Peter H Sims, Judd D Bowman, Steven G Murray, John P Barrett, Rigel C Cappallo, Colin J Lonsdale, Nivedita Mahesh, Raul A Monsalve, Alan E E Rogers, Titu Samson, and Akshatha K Vydula. A general bayesian model-validation framework based on null-test evidence ratios, with an example application to global 21-cm cosmology. *Monthly Notices of the Royal Astronomical Society*, 541(3):2262–2281, July 2025. ISSN 1365-2966. doi: 10.1093/mnras/staf1109. URL http://dx.doi.org/10.1093/mnras/staf1109.

[29] Peter H Sims, Judd D Bowman, Steven G Murray, John P Barrett, Rigel C Cappallo, Colin J Lonsdale, Nivedita Mahesh, Raul A Monsalve, Alan E E Rogers, Titu Samson, and Akshatha K Vydula. A bayesian approach to modelling spectrometer data chromaticity corrected using beam factors – II. model priors and posterior odds. *Monthly Notices of the Royal Astronomical Society*, 544(2):2340–2364, 12 2025. ISSN 0035-8711. doi: 10.1093/mnras/staf1767. URL https://doi.org/10.1093/mnras/staf1767.

[30] Jacob L Tutt, Peter H Sims, Joe H N Pattison, Dominic J Anstey, Samuel A K Leeney, and Eloy de Lera Acedo. Optimizing foreground modelling for global 21-cm cosmology with GPU-accelerated nested sampling. *Monthly Notices of the Royal Astronomical Society*, 550(1):stag1101, June 2026. ISSN 1365-2966. doi: 10.1093/mnras/stag1101. URL http://dx.doi.org/10.1093/mnras/stag1101.

[31] Anastasia Fialkov, Rennan Barkana, and Eli Visbal. The observable signature of late heating of the Universe during cosmic reionization. *Nature*, 506(7487):197–199, February 2014. doi: 10.1038/nature12999.

[32] P. A. Shaver, R. A. Windhorst, P. Madau, and A. G. de Bruyn. Can the reionization epoch be detected as a global signature in the cosmic background? *Astronomy & Astrophysics*, 345:380–390, May 1999. doi: 10.48550/arXiv.astro-ph/9901320.

[33] Dominic Anstey, Eloy de Lera Acedo, and Will Handley. A general bayesian framework for foreground modelling and chromaticity correction for global 21 cm experiments. *Monthly Notices of the Royal Astronomical Society*, 506(2):2041–2058, 06 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab1765. URL https://doi.org/10.1093/mnras/stab1765.

[34] Joshua J. Hibbard, Keith Tauscher, David Rapetti, and Jack O. Burns. Modeling the galactic foreground and beam chromaticity for global 21 cm cosmology. *The Astrophysical Journal*, 905(2):113, December 2020. ISSN 1538-4357. doi: 10.3847/1538-4357/abc3c5. URL http://dx.doi.org/10.3847/1538-4357/abc3c5.

[35] Neil Bassett, David Rapetti, Keith Tauscher, Jack O. Burns, and Joshua J. Hibbard. Ensuring robustness in training-set-based global 21 cm cosmology analysis. *The Astrophysical Journal*, 908(2):189, February 2021. ISSN 1538-4357. doi: 10.3847/1538-4357/abdb29. URL http://dx.doi.org/10.3847/1538-4357/abdb29.

[36] John Cumner, Carla Pieterse, Dirk de Villiers, and Eloy de Lera Acedo. The effects of the antenna power pattern uncertainty within a global 21 cm experiment. *Monthly Notices of the Royal Astronomical Society*, 531(4):4734–4745, June 2024. doi: 10.1093/mnras/stae1475. URL https://doi.org/10.1093/mnras/stae1475.

[37] Joe H. N. Pattison, John M. Cumner, Dominic J. Anstey, Saurabh Pegwal, Wessel Croukamp, Dirk I. L. de Villiers, and Eloy de Lera Acedo. Impact of antenna structure and orientation on forward-modelled global 21 cm signal recovery. *Monthly Notices of the Royal Astronomical Society*, 550(2):stag1216, 2026. doi: 10.1093/mnras/stag1216. URL https://doi.org/10.1093/mnras/stag1216.

[38] Aviad Cohen, Anastasia Fialkov, Rennan Barkana, and Matan Lotem. Charting the parameter space of the global 21-cm signal. *Monthly Notices of the Royal Astronomical Society*, 472 (2):1915–1931, 08 2017. ISSN 0035-8711. doi: 10.1093/mnras/stx2065. URL https://doi.org/10.1093/mnras/stx2065.

[39] R. F. Harrington. *Field Computation by Moment Methods*. Wiley–IEEE Press, Hoboken, New Jersey, US, 1993.

[40] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. doi: 10.1073/pnas.1912789117. URL https://doi.org/10.1073/pnas.1912789117.

[41] David Yallup, Namu Kroupa, and Will Handley. Nested slice sampling: Vectorized nested sampling for GPU-accelerated inference. *Transactions on Machine Learning Research*, May 2026. URL https://openreview.net/forum?id=5mF2eRl3gt.

[42] Alberto Cabezas, Adrien Corenflos, Junpeng Lao, and Rémi Louf. Blackjax: Composable Bayesian inference in JAX, 2024.

[43] Itamar Reis, Anastasia Fialkov, and Rennan Barkana. High-redshift radio galaxies: a potential new source of 21-cm fluctuations. *Monthly Notices of the Royal Astronomical Society*, 499(4):5993–6008, 10 2020. doi: 10.1093/mnras/staa3091.

[44] H T J Bevins, W J Handley, A Fialkov, E de Lera Acedo, and K Javid. <scp>globalemu</scp>: a novel and robust approach for emulating the sky-averaged 21-cm signal from the cosmic dawn and epoch of reionization. *Monthly Notices of the Royal Astronomical Society*, 508 (2):2923–2936, September 2021. ISSN 1365-2966. doi: 10.1093/mnras/stab2737. URL http://dx.doi.org/10.1093/mnras/stab2737.

[45] H T J Bevins, T Gessey-Jones, and W J Handley. On the accuracy of posterior recovery with neural network emulators. *Monthly Notices of the Royal Astronomical Society*, 544(1):375–390, September 2025. ISSN 1365-2966. doi: 10.1093/mnras/staf1590. URL http://dx.doi.org/10.1093/mnras/staf1590.

[46] Angélica de Oliveira-Costa, Max Tegmark, B. M. Gaensler, Justin Jonas, T. L. Landecker, and Patricia Reich. A model of diffuse Galactic radio emission from 10 MHz to 100 GHz. *Monthly Notices of the Royal Astronomical Society*, 388(1):247–260, July 2008. doi: 10.1111/j.1365-2966.2008.13376.x.

[47] Michael Pagano, Peter Sims, Adrian Liu, Dominic Anstey, Will Handley, and Eloy de Lera Acedo. A general bayesian framework to account for foreground map errors in global 21-cm experiments. *Monthly Notices of the Royal Astronomical Society*, 527(3):5649–5667, November 2023. ISSN 1365-2966. doi: 10.1093/mnras/stad3392. URL http://dx.doi.org/10.1093/mnras/stad3392.

[48] Jean Cavillot, Christophe Craeye, Eloy de Lera Acedo, and Nima Razavi-Ghods. Efficient MoM simulation of 3D metallic antenna connected to finite ground plane. In *2021 IEEE*

International Conference on Microwaves, Antennas, Communications and Electronic Systems (COMCAS), pages 213–216, 2021. doi: 10.1109/COMCAS52219.2021.9629009. URL https://doi.org/10.1109/COMCAS52219.2021.9629009.

[49] B. Hu, W. C. Chew, E. Michielssen, and J. Zhao. Fast inhomogeneous plane wave algorithm for the fast analysis of two-dimensional scattering problems. *Radio Science*, 34(4):759–772, 1999. doi: 10.1029/1999RS900038.

[50] K. Alkhalifeh, G. Hislop, N. A. Ozdemir, and C. Craeye. Efficient mom simulation of 3-d antennas in the vicinity of the ground. *IEEE Transactions on Antennas and Propagation*, 64 (12):5335–5344, 2016. doi: 10.1109/TAP.2016.2618482.

[51] Jean Cavillot, Denis Tihon, Francisco Mesa, Eloy de Lera Acedo, and Christophe Craeye. Efficient simulation of large irregular arrays on a finite ground plane. *IEEE Transactions on Antennas and Propagation*, 68:2753–2764, 2020. doi: 10.1109/TAP.2019.2955180.

[52] Jean Cavillot, Denis Tihon, Quentin Gueuning, Eloy de Lera Acedo, and Christophe Craeye. Full-wave analysis of thermal noise in antenna arrays on top of layered medium. *IEEE Transactions on Antennas and Propagation*, 72:7560–7573, 2024. doi: 10.1109/TAP.2024.3445132.

[53] Jacob L. Tutt, Dominic J. Anstey, John Cumner, Harry T. J. Bevins, Jean Cavillot, and Eloy de Lera Acedo. Towards end-to-end bayesian forward models in global 21-cm cosmology: surrogate modelling and marginalisation of beam uncertainty, 2026. URL https://arxiv.org/abs/2608.18962.

[54] Carla M. Pieterse, Mariet Venter, and Dirk I. L. de Villiers. Investigating frequency dependency in characteristic basis function pattern modelling with geometric perturbations. In *2024 International Conference on Electromagnetics in Advanced Applications (ICEAA)*, pages 407–412, 2024. doi: 10.1109/ICEAA61917.2024.10701964. URL https://doi.org/10.1109/ICEAA61917.2024.10701964.

[55] K. H. Scheutwinkel, E. de Lera Acedo, and W. Handley. Bayesian evidence-driven diagnosis of instrumental systematics for sky-averaged 21-cm cosmology experiments. *Publications of the Astronomical Society of Australia*, 39:e052, 2022. ISSN 1448-6083. doi: 10.1017/pasa.2022.49. URL http://dx.doi.org/10.1017/pasa.2022.49.

[56] Shikhar Mittal, Girish Kulkarni, Dominic Anstey, and Eloy de Lera Acedo. Impact of extragalactic point sources on the low-frequency sky spectrum and cosmic dawn global 21-cm measurements. *Monthly Notices of the Royal Astronomical Society*, 534(2):1317–1328, 09 2024. ISSN 0035-8711. doi: 10.1093/mnras/stae2111. URL https://doi.org/10.1093/mnras/stae2111.

[57] Daniel Robins, Dominic Anstey, Harry Bevins, Eloy de Lera Acedo, and Melis O. Irfan. Synchrotron and free–free mapping with simulated REACH observations between 50 and 170 MHz. *Monthly Notices of the Royal Astronomical Society*, 552(4):stag1804, 2026. doi: 10.1093/mnras/stag1804. URL https://doi.org/10.1093/mnras/stag1804.

[58] Emma Shen, Dominic Anstey, Eloy de Lera Acedo, Anastasia Fialkov, and Will Handley. Quantifying ionospheric effects on global 21-cm observations. *Monthly Notices of the Royal Astronomical Society*, 503(1):344–353, 02 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab429. URL https://doi.org/10.1093/mnras/stab429.

[59] Emma Shen, Dominic Anstey, Eloy de Lera Acedo, and Anastasia Fialkov. Bayesian data analysis for sky-averaged 21-cm experiments in the presence of ionospheric effects. *Monthly Notices of the Royal Astronomical Society*, 515(3):4565–4573, 07 2022. ISSN 0035-8711. doi: 10.1093/mnras/stac1900. URL https://doi.org/10.1093/mnras/stac1900.

[60] Joe H N Pattison, Dominic J Anstey, and Eloy de Lera Acedo. Modelling a hot horizon in global 21-cm experimental foregrounds. *Monthly Notices of the Royal Astronomical Society*, 527(2):2413–2425, 11 2023. ISSN 0035-8711. doi: 10.1093/mnras/stad3378. URL https://doi.org/10.1093/mnras/stad3378.

[61] Joe H N Pattison, Jean Cavillot, Harry T J Bevins, Dominic J Anstey, John M Cumner, and Eloy de Lera Acedo. Global 21 cm signal recovery under changing environmental conditions. *Monthly Notices of the Royal Astronomical Society*, 538(3):1301–1313, 02 2025. ISSN 0035-8711. doi: 10.1093/mnras/staf315. URL https://doi.org/10.1093/mnras/staf315.

[62] Joe H. N. Pattison, Dominic J. Anstey, and Eloy de Lera Acedo. Quantifying the impact of lunar and planetary occultation on experimental global 21 cm cosmology. *Monthly Notices of the Royal Astronomical Society*, 546(4):stag258, 2026. doi: 10.1093/mnras/stag258.

[63] Ian L. V. Roque et al. Receiver design for the REACH global 21-cm signal experiment. *Experimental Astronomy*, 59(1):7, 2025. doi: 10.1007/s10686-024-09975-3. URL https://doi.org/10.1007/s10686-024-09975-3.

[64] Christian J. Kirkham, William J. Handley, Jiacong Zhu, Kaan Artuc, Ian L. V. Roque, Samuel A. K. Leeney, Harry T. J. Bevins, Dominic J. Anstey, and Eloy de Lera Acedo. Accounting for noise and singularities in bayesian calibration methods for global 21-cm cosmology experiments. *Monthly Notices of the Royal Astronomical Society*, 543(4):4312–4320, 2025. doi: 10.1093/mnras/staf1715. URL https://doi.org/10.1093/mnras/staf1715.

[65] W. J. Handley, M. P. Hobson, and A. N. Lasenby. polychord: next-generation nested sampling. *Monthly Notices of the Royal Astronomical Society*, 453(4):4384–4398, 09 2015. ISSN 0035-8711. doi: 10.1093/mnras/stv1911. URL https://doi.org/10.1093/mnras/stv1911.

[66] Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 343–351. PMLR, 2021. URL https://proceedings.mlr.press/v130/lueckmann21a.html.

[67] James Alvey, Uddipta Bhardwaj, Valerie Domcke, Mauro Pieroni, and Christoph Weniger. Simulation-based inference for stochastic gravitational wave background data analysis. *Physical Review D*, 109(8):083008, 2024. doi: 10.1103/PhysRevD.109.083008. URL https://doi.org/10.1103/PhysRevD.109.083008.

[68] Pedro J. Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F. Podlaski, Sara A. Haddad, Tim P. Vogels, David S. Greenberg, and Jakob H. Macke. Training deep neural density estimators to identify mechanistic models of neural dynamics. *eLife*, 9:e56261, 2020. doi: 10.7554/eLife.56261. URL https://doi.org/10.7554/eLife.56261.

[69] Michael Deistler, Kyra L. Kadhim, Matthijs Pals, Jonas Beck, Ziwei Huang, Manuel Gloeckler, Janne K. Lappalainen, Cornelius Schr"oder, Philipp Berens, Pedro J. Gonçalves, and Jakob H. Macke. Jaxley: differentiable simulation enables large-scale training of detailed biophysical models of neural dynamics. *Nature Methods*, 22:2649–2657, 2025. doi: 10.1038/s41592-025-02895-w. URL https://doi.org/10.1038/s41592-025-02895-w.

[70] Ian Christopher Tanoh, Michael Deistler, Jakob H. Macke, and Scott W. Linderman. Identifying multi-compartment Hodgkin–Huxley models with high-density extracellular voltage recordings. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3307. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8e9c9bbe03d0da2d61e648e751ae61a4-Abstract-Conference.html.

[71] David Yallup. Nested sampling with slice-within-Gibbs: Efficient evidence calculation for hierarchical Bayesian models, 2026. URL https://arxiv.org/abs/2602.17414.

[72] David Yallup, Metha Prathaban, James Alvey, and Will Handley. Parallel nested slice sampling for gravitational wave parameter estimation, 2025. URL https://arxiv.org/abs/2509.24949.

[73] Michael Deistler, Jan Boelts, Peter Steinbach, Guy Moss, Thomas Moreau, Manuel Gl"ockler, Pedro L. C. Rodrigues, Julia Linhart, Janne K. Lappalainen, Benjamin Kurt Miller, Pedro J. Gonçalves, Jan-Matthis Lueckmann, Cornelius Schr"oder, and Jakob H. Macke. Simulation-based inference: A practical guide, 2025. URL https://arxiv.org/abs/2508.12939.

[74] Manuel Glöckler, Michael Deistler, and Jakob H. Macke. Variational methods for simulation-based inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=kZ0UYdhqkNY.

[75] Huifang Lyu, James Alvey, Noemi Anau Montel, Mauro Pieroni, and Christoph Weniger. Dynamic SBI: round-free sequential simulation-based inference with adaptive datasets. *Machine Learning: Science and Technology*, 7(2):025036, 2026. doi: 10.1088/2632-2153/ae51e1. URL https://doi.org/10.1088/2632-2153/ae51e1.

[76] James Alvey, Carlo R. Contaldi, and Mauro Pieroni. Simulation-based inference with deep ensembles: evaluating calibration uncertainty and detecting model misspecification. *Machine Learning: Science and Technology*, 7(1):015008, 2026. doi: 10.1088/2632-2153/ae3103. URL https://doi.org/10.1088/2632-2153/ae3103.

[77] S. A. K. Leeney, T. Gessey-Jones, W. J. Handley, E. de Lera Acedo, H. T. J. Bevins, and J. L. Tutt. Conditional neural bayes ratio estimation for experimental design optimisation, 2026. URL https://arxiv.org/abs/2603.26489.

[78] Diana Cai, Chirag Modi, Charles C. Margossian, Robert M. Gower, David M. Blei, and Lawrence K. Saul. EigenVI: score-based variational inference with orthogonal function expansions. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4218. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ef72fa6579401ffff9da246a5014f055-Abstract-Conference.html.

[79] Geetakrishnasai Gunapati, Anirudh Jain, P. K. Srijith, and Shantanu Desai. Variational inference as an alternative to MCMC for parameter estimation and model selection. *Publications of the Astronomical Society of Australia*, 39:e001, 2022. ISSN 1448-6083. doi: 10.1017/pasa.2021.64. URL http://dx.doi.org/10.1017/pasa.2021.64.

[80] Simon McIntosh-Smith, Sadaf R. Alam, and Christopher J. Woods. Isambard-AI: A leadership-class supercomputer optimised specifically for artificial intelligence. In *CUG 2024 Proceedings*, pages 44–54, 2024.

[81] Simon Pochinda, Thomas Gessey-Jones, Harry T. J. Bevins, Anastasia Fialkov, Stefan Heimersheim, Irene Abril-Cabezas, Eloy de Lera Acedo, Saurabh Singh, Shikhar Sikder, and Rennan Barkana. Constraining the properties of population iii galaxies with multi-wavelength observations. *Monthly Notices of the Royal Astronomical Society*, 531(1):1113–1132, 2024. doi: 10.1093/mnras/stae1185.

[82] Jiten Dhandha, Anastasia Fialkov, Thomas Gessey-Jones, Harry T. J. Bevins, Sandro Tacchella, Simon Pochinda, Eloy de Lera Acedo, Saurabh Singh, and Rennan Barkana. Exploiting synergies between JWST and cosmic 21-cm observations to uncover star formation in the early universe. *Monthly Notices of the Royal Astronomical Society*, 542(3):2292–2322, 2025. doi: 10.1093/mnras/staf1359. URL https://doi.org/10.1093/mnras/staf1359.

[83] HERA Collaboration. HERA phase I limits on the cosmic 21-cm signal: Constraints on astrophysics and cosmology during the epoch of reionization. *The Astrophysical Journal*, 924 (2):51, 2022. doi: 10.3847/1538-4357/ac2ffc.

[84] HERA Collaboration. First results from HERA phase II. *The Astrophysical Journal*, 998(1):33, 2026. doi: 10.3847/1538-4357/ae2d54.

[85] HERA Collaboration. Improved constraints on the 21 cm EoR power spectrum and the x-ray heating of the IGM with HERA phase I observations. *The Astrophysical Journal*, 945(2):124, 2023. doi: 10.3847/1538-4357/acaf50.

[86] John Skilling. Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4): 833–859, 2006. doi: 10.1214/06-BA127.

A Additional implementation and experimental details

A.1 Astrophysical signal model

`21cmSPACE` is a semi-numerical model that evolves a comoving simulation volume divided into 128^3 cells, each with a volume of $(3\ \mathrm{cMpc})^3$, from $z = 50$ to the end of reionisation at $z = 6$ [31, 38]. Large-scale density and radiation fields are evolved explicitly, while halo and star formation are treated through sub-grid prescriptions. The resulting stellar populations source the Lyα, X-ray, and ionising radiation that determines the coupling, temperature, and neutral fraction of the IGM, from which the sky-averaged brightness temperature $T_{21}(z)$ is calculated.

We adopt the inhomogeneous radio-galaxy extension introduced by Reis et al. [43] and subsequently used in analyses of the first galaxies and HERA power-spectrum constraints [81–84]. High-redshift galaxies are assigned a radio luminosity proportional to their star-formation rate,

$$\frac{L_{\mathrm{r}}(\nu)}{\mathrm{W\,Hz^{-1}}} = f_{\mathrm{radio}} 10^{22} \left(\frac{\nu}{150\ \mathrm{MHz}}\right)^{-0.7} \left(\frac{\mathrm{SFR}}{M_\odot\ \mathrm{yr^{-1}}}\right), \tag{6}$$

where $f_{\mathrm{radio}} = 1$ corresponds to the luminosity of present-day radio galaxies. `21cmSPACE` propagates this spatially varying emission, modifying the radiation temperature against which the 21-cm transition is observed. Table 1 lists the nine quantities over which the simulation suite and emulator are defined.

Table 1: Physical parameters and training ranges of the `21cmSPACE` simulations. The parameters $f_{\star,\mathrm{II}}$, $f_{\star,\mathrm{III}}$, V_{c}, f_{X}, and f_{radio} are sampled uniformly in their base-10 logarithm, τ is sampled uniformly in linear space, and α, pop, and E_{min} are evaluated on discrete grids.

Parameter	Physical role	Training range
$f_{\star,\mathrm{II}}$	Pop II star-formation efficiency	10^{-3}–$10^{-0.3}$
$f_{\star,\mathrm{III}}$	Pop III star-formation efficiency	10^{-3}–$10^{-0.3}$
V_{c}	Minimum halo circular velocity for star formation	4.2–100 $\mathrm{km\,s^{-1}}$
f_{X}	X-ray luminosity efficiency	10^{-3}–10^{3}
τ	Thomson optical depth	0.0235–0.1021
f_{radio}	Radio luminosity relative to present-day galaxies	10^{-1}–10^{5}
α	Power-law slope of the X-ray SED	$\{1.0, 1.3, 1.5\}$
pop	Discrete Pop III stellar-population prescription	$\{231, 232, 233\}$
E_{min}	Minimum escaping X-ray energy	$\{100, 200, \ldots, 1500, 2000, 3000\}$ eV

A.2 Neural-emulator dataset, training, and validation

The simulation ensemble contains 109,525 complete $T_{21}(z)$ realisations spanning the parameter space in Table 1.[6] Each simulation is evaluated on 200 redshift samples over $6 \leq z \leq 27$. We divide the complete simulations into training, validation, and test sets in a 60:20:20 ratio. This split is performed at the simulation level before redshift is folded into the network input, preventing samples from the same physical realisation from appearing in more than one subset.

Following the `globalemu` formulation [44], redshift is treated as an additional input and the network learns the scalar mapping

$$\widehat{T}_{21}(z; \boldsymbol{\Theta}) = f_{\boldsymbol{\phi}}(z, \boldsymbol{\Theta}), \qquad \boldsymbol{\Theta} = \{f_{\star,\mathrm{II}}, f_{\star,\mathrm{III}}, V_{\mathrm{c}}, f_{\mathrm{X}}, \alpha, E_{\mathrm{min}}, \tau, f_{\mathrm{radio}}, \mathrm{pop}\}. \tag{7}$$

The scale-spanning parameters $f_{\star,\mathrm{II}}$, $f_{\star,\mathrm{III}}$, V_{c}, f_{X}, and f_{radio} are transformed to base-10 logarithmic coordinates, while α, E_{min}, and pop enter as discrete conditioning variables drawn from their simulated grids. All inputs are standardised using the training set.

The emulator is a dense multilayer perceptron with four hidden layers of 64 units and GELU activations. We train against the simulated temperature in linear mK using a mean-squared-error objective and AdamW with a batch size of 1,000 and weight decay of 10^{-4}. Training comprises 2,000 epochs: the learning rate decays exponentially from 10^{-3} to 5×10^{-5} over the first 1,000 epochs, then from 5×10^{-5} to 10^{-5} during a 1,000-epoch continuation. Early stopping is triggered

[6] These simulations were precomputed for previous analyses [81–85], at a total computational cost of $\sim 9.20 \times 10^5$ CPU-core hours on Intel Xeon Platinum 8368Q processors with 6.76 GB of RAM per core.

at epoch 1,943 and this checkpoint is retained.[7] Across the 21,905 held-out test spectra, the emulator achieves a mean per-signal RMSE of $0.91\ \mathrm{mK}$.

For the benchmark presented here, we condition the emulator on $\alpha = 1.3$, $E_{\min} = 500\ \mathrm{eV}$, and $\mathrm{pop} = 232$. The directly inferred signal space is therefore

$$\boldsymbol{\theta}_{21} = \{\log_{10} f_{\star,\mathrm{II}}, \log_{10} f_{\star,\mathrm{III}}, \log_{10} V_{\mathrm{c}}, \log_{10} f_{\mathrm{X}}, \tau, \log_{10} f_{\mathrm{radio}}\}. \tag{8}$$

Uniform priors over the emulator-training coordinates correspond to log-uniform priors on $f_{\star,\mathrm{II}}$, $f_{\star,\mathrm{III}}$, V_{c}, f_{X}, and f_{radio}, and a uniform prior on τ in linear space.

A.3 Foreground construction and priors

To construct a full-fidelity foreground template that can be extrapolated across the observing band, we follow Anstey et al. [33] and evaluate the 2008 Global Sky Model (GSM) [46] at 230 and 408 MHz. In celestial coordinates Ω', these maps define the spatially varying spectral-index field

$$\beta_{\mathrm{GSM}}(\Omega') = -\frac{\log\big[\big(T_{230}(\Omega') - T_{\mathrm{CMB}}\big)/\big(T_{408}(\Omega') - T_{\mathrm{CMB}}\big)\big]}{\log(230/408)}, \tag{9}$$

where $T_{\mathrm{CMB}} = 2.725\ \mathrm{K}$. This field is then used to extrapolate the 230 MHz template to an arbitrary observing frequency:

$$T_{\mathrm{sky}}(\Omega', \nu) = [T_{230}(\Omega') - T_{\mathrm{CMB}}] \left(\frac{\nu}{230\ \mathrm{MHz}}\right)^{-\beta_{\mathrm{GSM}}(\Omega')} + T_{\mathrm{CMB}}. \tag{10}$$

This continuous pixel-level model is used to generate the benchmark data described in Appendix A.5.

For inference, the spatial distributions of both the base-map temperature and spectral index are uncertain and cannot be treated as exact templates. A full pixel-level parameterisation would, however, be highly degenerate for the one- or two-dimensional data produced by a single radiometer. We therefore follow the regional construction of Anstey et al. [23, 33] and Pagano et al. [47], partitioning the sky independently into regions of uniform spectral index and base-map amplitude. In this work, we define $N_\beta = 30$ spectral-index regions using equal percentiles of the full-resolution β_{GSM} field, such that each binary mask $M_{\beta,j}(\Omega')$ contains an equal fraction of the sky pixels. An equivalent set of masks $M_{\alpha,i}(\Omega')$ defines regions of uniform base-map correction. Their intersections give the two-dimensional masks used in Equation 1, $M_{i,j}(\Omega') = M_{\alpha,i}(\Omega')M_{\beta,j}(\Omega')$.

Each regional spectral-index parameter is assigned the same uniform prior spanning the minimum and maximum of the full spectral-index field,

$$\beta_j \sim \mathcal{U}(\beta_{\min}, \beta_{\max}), \qquad \beta_{\min} = \min_{\Omega'} \beta_{\mathrm{GSM}}(\Omega'), \quad \beta_{\max} = \max_{\Omega'} \beta_{\mathrm{GSM}}(\Omega'). \tag{11}$$

If uncertainty in the base-map amplitude is included, the corresponding multiplicative corrections are assigned

$$\alpha_i \sim \mathcal{N}(1, 0.2^2), \tag{12}$$

representing a 20% uncertainty about the GSM template. The benchmark in this work isolates spectral-index and beam uncertainty and therefore uses a single amplitude region, $N_\alpha = 1$, fixed to $\alpha_1 = 1$. The foreground consequently contributes 30 parameters to the sampled space.

A.4 Beam-surrogate construction and validation

We characterise each simulated beam using its normalised directivity, $D(\nu, \Omega)$, calculated from the orthogonal far-field electric-field components as

$$D(\nu, \Omega) = 4\pi \frac{|E_\theta(\nu, \Omega)|^2 + |E_\phi(\nu, \Omega)|^2}{\displaystyle\int_{4\pi} \big[|E_\theta(\nu, \Omega)|^2 + |E_\phi(\nu, \Omega)|^2\big]\ \mathrm{d}\Omega}. \tag{13}$$

The below-horizon response is set to zero before each frequency channel is renormalised such that $\int_{4\pi} D(\nu_i, \Omega)\,\mathrm{d}\Omega = 4\pi$. We generate 500 beam realisations using the Method-of-Moments CEM solver described by Cavillot et al. [48, 51, 52]. Latin-hypercube sampling distributes these simulations

[7] Training required approximately 73 minutes on a single NVIDIA H100 GPU.

across the seven physical perturbations listed in Table 2. The ensemble is randomly divided into 400 training beams and 100 held-out validation beams.[8]

Table 2: Physical parameters and ranges used to generate the CEM beam ensemble. Rotations and displacements are defined relative to the ground plane.

Parameter	Symbol	Range
Dipole blade height	H_{blade}	800–1200 mm
Ground-plane height	H_{GP}	1000–1500 mm
Dipole rotation	α_{rel}	-10–10 deg
Dipole x displacement	Δx_{rel}	-200–200 mm
Dipole y displacement	Δy_{rel}	-200–200 mm
Soil permittivity, real	ϵ_{real}	1–10
Soil permittivity, imaginary	ϵ_{imag}	-2–0

The first stage of the surrogate follows Pieterse et al. [54]. For each training beam b and frequency channel ν_i, the directivity is flattened over its N_Ω angular samples and stored as one row of the matrix $\mathbf{M} \in \mathbb{R}^{(N_{\mathrm{train}}N_\nu)\times N_\Omega}$. Writing the combined beam–frequency row index as $\gamma = (b, i)$, its singular value decomposition is

$$M_{\gamma p} = \sum_{k=1}^{N_{\mathrm{mode}}} U_{\gamma k}\, \sigma_k\, (V^\dagger)_{kp}, \tag{14}$$

where p indexes angular samples, $N_{\mathrm{mode}} = \min(N_{\mathrm{train}}N_\nu, N_\Omega)$, and the singular values σ_k are ordered by decreasing contribution to the beam variance. The rows of $\mathbf{V}^\dagger$ define the angular basis functions $B_k(\Omega)$, while the beam- and frequency-dependent weights are

$$w_{b,k}(\nu_i) = \sigma_k U_{(b,i),k}. \tag{15}$$

A beam reconstructed with the first R modes is therefore

$$\widetilde{D}_b^{(R)}(\nu_i, \Omega) = \sum_{k=1}^{R} w_{b,k}(\nu_i) B_k(\Omega). \tag{16}$$

The truncation rank is selected in antenna-temperature space, rather than from directivity error alone. Each truncated reconstruction is convolved with the full-resolution foreground sky for a conservative Galactic-centre-overhead snapshot,[9] and compared with the antenna temperature obtained using the corresponding full CEM beam. Retaining $R = 400$ angular modes gives a mean RMSE of 20.09 mK across the training ensemble, below the adopted 25 mK instrumental-noise level. Directly parameterising the resulting frequency-dependent weights would nevertheless require $RN_\nu = 34{,}000$ coefficients and would not encode their correlations across frequency and angular mode.

We therefore perform a second decomposition in the SVD-weight space. For each beam, the retained weights are stacked over frequency and angular mode into a vector $\mathbf{w}_b \in \mathbb{R}^{RN_\nu}$. After subtracting the training-set mean $\mathbf{e}_0$, an SVD of the residual beam-to-beam variations yields a new orthogonal set of correction vectors $\{\mathbf{e}_m\}$. In component form, the reconstructed weights are

$$\widetilde{w}_{b,k}^{(M)}(\nu_i) = e_{0,k}(\nu_i) + \sum_{m=1}^{M} a_{b,m} e_{m,k}(\nu_i), \tag{17}$$

which, when substituted into Equation 16, gives the surrogate in Equation 2. This second stage preserves the coherent variation of the beam across the observing band while reducing the instrumental representation to the coefficient vector $\boldsymbol{a}$.

The beam coefficients are centred by construction and assigned the Gaussian approximation

$$\boldsymbol{a} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_a), \qquad \boldsymbol{\Sigma}_a = \mathrm{diag}[\mathrm{Var}(a_{b,1}), \ldots, \mathrm{Var}(a_{b,M})]\,, \tag{18}$$

[8] Generating the complete ensemble required 8.0×10^5 CPU-core hours on Intel Xeon Platinum 8368Q processors with 6.76 GB of RAM per core.

[9] Evaluated at 00:00:00 UTC on 2019-07-01 at the observing site (lat $= -30.83875°$, lon $= 21.37492°$, altitude 1151 m), corresponding to an apparent LST of 20:00:31.86.

where the variances are estimated across the 400 training beams. Applying the same antenna-temperature criterion after both decomposition stages shows that $M = 100$ correction modes give a mean reconstruction RMSE of $24.47\ \mathrm{mK}$. The final surrogate therefore compresses the original 34,000-dimensional SVD-weight space by a factor of 340. All benchmark observations use a full CEM beam from the held-out validation set, rather than a reconstruction used to train the surrogate (see Tutt et al. [53] for details).

A.5 Benchmark data generation

The benchmark datasets are generated independently of the hybrid representations used during inference. At each observing time, the full-resolution sky in Equation 10 is rotated from Ω' into the local antenna frame Ω and coupled to a full CEM validation beam, $D_{\mathrm{val}}(\nu, \Omega)$, which was excluded from the surrogate training set. For injected signal realisation s, the simulated data are

$$d_s(\nu, t) = \frac{1}{4\pi} \int_{4\pi} D_{\mathrm{val}}(\nu, \Omega) T_{\mathrm{sky}}(\Omega, \nu, t)\, \mathrm{d}\Omega + T_{21}(\nu; \boldsymbol{\theta}_{21,s}) + \epsilon(\nu, t), \tag{19}$$

where $\epsilon(\nu, t) \sim \mathcal{N}(0, \sigma_n^2)$ is uncorrelated Gaussian noise with $\sigma_n = 25\ \mathrm{mK}$.

We consider a three-hour observation from an observational site in the Karoo Desert, South Africa, beginning at 00:00:00 UTC on 2019-10-01. The observation is discretised into 1 MHz channels from 50 to 135 MHz and 5-minute time bins, giving $N_\nu = 86$ and $N_t = 37$. A test beam selected from the held-out set is used throughout, with $\{H_{\mathrm{blade}}, H_{\mathrm{GP}}, \alpha_{\mathrm{rel}}, \Delta x_{\mathrm{rel}}, \Delta y_{\mathrm{rel}}, \epsilon_{\mathrm{real}}, \epsilon_{\mathrm{imag}}\} = \{1080\ \mathrm{mm}, 1272\ \mathrm{mm}, 9.8^\circ, 18\ \mathrm{mm}, -91\ \mathrm{mm}, 4.24, -0.68\}$. The full-resolution foreground, beam, and observing window are fixed across the benchmark, isolating performance as the astrophysical signal changes. For further work investigating the robustness of the inference framework to different observing windows and beam relaisations see Tutt et al. [30] and Tutt et al. [53] respecitvely.

The 100 injected profiles are independent random draws from the six-dimensional prior defined in Appendix A.1. Consequently, the suite samples both strongly featured signals and broad, weakly identifiable profiles from the stated astrophysical model.

A.6 Analytical beam marginalisation

For the marginalisation, the frequency–time data grid is flattened into a vector $\mathbf{d} \in \mathbb{R}^{N_{\mathrm{data}}}$, where $N_{\mathrm{data}} = N_\nu N_t$. With the amplitude corrections fixed to $\alpha_i = 1$, we collect the sampled parameters as $\boldsymbol{\phi} = (\boldsymbol{\theta}_{\mathrm{FG}}, \boldsymbol{\theta}_{21}, \boldsymbol{\theta}_{\mathrm{Noise}})$. The mean-beam model is

$$\mu(\nu, t; \boldsymbol{\phi}) = \sum_{j=1}^{N_\beta} \mathcal{B}_{1,j,0}(\nu, t) \left(\frac{\nu}{\nu_0}\right)^{-\beta_j} + T_{\mathrm{CMB}} + T_{21}(\nu; \boldsymbol{\theta}_{21}), \tag{20}$$

and the response of beam-correction mode m is

$$h_m(\nu, t; \boldsymbol{\beta}) = \sum_{j=1}^{N_\beta} \mathcal{B}_{1,j,m}(\nu, t) \left(\frac{\nu}{\nu_0}\right)^{-\beta_j}. \tag{21}$$

Evaluating these expressions over the complete grid gives the vector $\boldsymbol{\mu}$ and design matrix

$$\mathbf{H}(\boldsymbol{\beta}) = [\mathbf{h}_1\ \mathbf{h}_2\ \cdots\ \mathbf{h}_M], \tag{22}$$

such that the forward-modelled data are linear in the beam coefficients,

$$\mathbf{d} = \boldsymbol{\mu}(\boldsymbol{\phi}) + \mathbf{H}(\boldsymbol{\beta})\boldsymbol{a} + \boldsymbol{n}. \tag{23}$$

We assume the beam prior in Equation 18 and Gaussian noise $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_n(\theta_n))$. The likelihood conditional on the beam coefficients is therefore

$$p(\mathbf{d} \mid \boldsymbol{\phi}, \boldsymbol{a}) = \mathcal{N}[\mathbf{d} \,|\, \boldsymbol{\mu}(\boldsymbol{\phi}) + \mathbf{H}(\boldsymbol{\beta})\boldsymbol{a}, \mathbf{C}_n(\theta_n)]. \tag{24}$$

Because both this likelihood and the prior on $\boldsymbol{a}$ are Gaussian, the instrumental parameters can be integrated out exactly:

$$\begin{aligned} p(\mathbf{d} \mid \boldsymbol{\phi}) &= \int p(\mathbf{d} \mid \boldsymbol{\phi}, \boldsymbol{a}) p(\boldsymbol{a})\, \mathrm{d}\boldsymbol{a} \\ &= \mathcal{N}[\mathbf{d} \,|\, \boldsymbol{\mu}(\boldsymbol{\phi}), \mathbf{C}_{\mathrm{eff}}(\boldsymbol{\phi})], \end{aligned} \tag{25}$$

where

$$\mathbf{C}_{\text{eff}} = \mathbf{C}_n + \mathbf{H}\boldsymbol{\Sigma}_a\mathbf{H}^{\mathrm{T}}. \tag{26}$$

Defining the residual $\mathbf{r} = \mathbf{d} - \boldsymbol{\mu}$, the marginal log-likelihood is

$$\ln \mathcal{L}_{\text{marg}} = -\frac{1}{2}\left[\mathbf{r}^{\mathrm{T}}\mathbf{C}_{\text{eff}}^{-1}\mathbf{r} + \ln\det \mathbf{C}_{\text{eff}} + N_{\text{data}}\ln(2\pi)\right]. \tag{27}$$

Analytical marginalisation reduces the explicitly sampled model from 137 uncertain quantities to the 30 foreground, six astrophysical, and one noise parameter used by NSS, while retaining the covariance induced by the complete beam prior in both the posterior and evidence.

A.7 Nested Slice Sampling configuration

We use Nested Sampling (NS) [86] to obtain posterior samples and the Bayesian evidence,

$$\mathcal{Z} = \int \mathcal{L}(\boldsymbol{\phi})\pi(\boldsymbol{\phi})\,\mathrm{d}\boldsymbol{\phi} = \int_0^1 \mathcal{L}(X)\,\mathrm{d}X, \tag{28}$$

where X is the prior mass enclosed above a likelihood contour. Standard NS evolves a population of live points by repeatedly removing the lowest-likelihood point and replacing it with a draw satisfying the resulting likelihood constraint.

We use the GPU-native Nested Slice Sampling (NSS) implementation in `BlackJAX` [41, 42]. As implemented by Yallup et al. [41], each iteration removes the b lowest-likelihood points and evolves b independent replacement chains in parallel, all constrained above the highest likelihood among the removed batch. The expected contraction in log prior mass is

$$\mathbb{E}[\Delta \log X] = -\sum_{i=0}^{b-1}\frac{1}{n_{\text{live}} - i} \simeq \log\left(\frac{n_{\text{live}} - b}{n_{\text{live}}}\right). \tag{29}$$

Each replacement begins from a randomly selected surviving live point and is evolved using hit-and-run slice sampling within the likelihood-constrained region. Proposal directions are drawn in the Mahalanobis-normalised geometry defined by the covariance of the live population; stepping out and interval shrinkage then identify a valid point along each direction.

After analytical beam marginalisation, the sampled space has $D = 37$ dimensions. All benchmark runs use $n_{\text{live}} = 25D = 925$ live points, $n_{\text{inner}} = 8D = 296$ slice steps per replacement, and $b = 0.2n_{\text{live}} = 185$ simultaneous replacements. The live population is initialised independently from the prior for each injected dataset, and sampling terminates when the estimated remaining live-point evidence satisfies $\log \mathcal{Z}_{\text{live}} - \log \mathcal{Z} < -3$.

A.8 Signal-profile recovery diagnostics

Recovery is evaluated through complementary diagnostics of posterior-mean accuracy and posterior calibration. For each benchmark realisation r, we systematically resample $S = 2048$ equal-weight profiles, $T_r^{(s)}(\nu_k)$, from the weighted NSS posterior over K frequency channels. For injected profile $T_r^{\star}(\nu_k)$, the posterior moments and RMSE are

$$\begin{aligned}
\overline{T}_{r,k} &= \frac{1}{S}\sum_{s=1}^{S} T_r^{(s)}(\nu_k), & \widehat{\sigma}_{r,k} &= \left[\frac{1}{S-1}\sum_{s=1}^{S}\left(T_r^{(s)}(\nu_k) - \overline{T}_{r,k}\right)^2\right]^{1/2}, \\
\mathrm{RMSE}_r &= \left[\frac{1}{K}\sum_{k=1}^{K}\left(T_r^{\star}(\nu_k) - \overline{T}_{r,k}\right)^2\right]^{1/2}.
\end{aligned} \tag{30}$$

RMSE quantifies point accuracy but not uncertainty calibration. We therefore assess coverage using the maximum marginally standardised deviation across frequency, a stringent whole-profile criterion requiring the credible band to contain every frequency channel:

$$\begin{aligned}
Z_r^{\star} &= \max_k \left|\frac{T_r^{\star}(\nu_k) - \overline{T}_{r,k}}{\widehat{\sigma}_{r,k}}\right|, & Z_r^{(s)} &= \max_k \left|\frac{T_r^{(s)}(\nu_k) - \overline{T}_{r,k}}{\widehat{\sigma}_{r,k}}\right|, \\
\mathcal{C}_{r,c}(\nu_k) &= \left[\overline{T}_{r,k} - q_{r,c}\widehat{\sigma}_{r,k}, \overline{T}_{r,k} + q_{r,c}\widehat{\sigma}_{r,k}\right].
\end{aligned} \tag{31}$$

where $q_{r,c}$ is the empirical c-quantile of $\{Z_r^{(s)}\}_{s=1}^S$. Estimating this quantile from complete posterior profiles retains their dependence across frequency. The minimum credible level required to contain the injection and the empirical coverage over R realisations are

$$\begin{aligned} \widehat{c}_r &= \frac{1}{S}\sum_{s=1}^{S}\left[\mathbb{I}\Big(Z_r^{(s)} < Z_r^{\star}\Big) + \frac{1}{2}\mathbb{I}\Big(Z_r^{(s)} = Z_r^{\star}\Big)\right], \\ \widehat{\mathrm{Cov}}(c) &= \frac{1}{R}\sum_{r=1}^{R}\mathbb{I}(\widehat{c}_r \le c)\,. \end{aligned} \tag{32}$$

For $R = 100$, $\widehat{\mathrm{Cov}}(0.95) = 0.94$, consistent with the nominal 95% level. Figure 3 is ordered by $\widehat{c}_r$ but displays pointwise intervals; coverage uses the whole-profile bands in Equation 31.

A.9 Astrophysical parameter posteriors

To complement the signal-space recovery analysis in Section 3.1, Figure 4 presents the corresponding six-dimensional astrophysical posteriors for the four representative cases shown in Figure 3. These distributions illustrate how intrinsic parameter degeneracies and multimodality can map distinct astrophysical scenarios to similar global signal profiles.

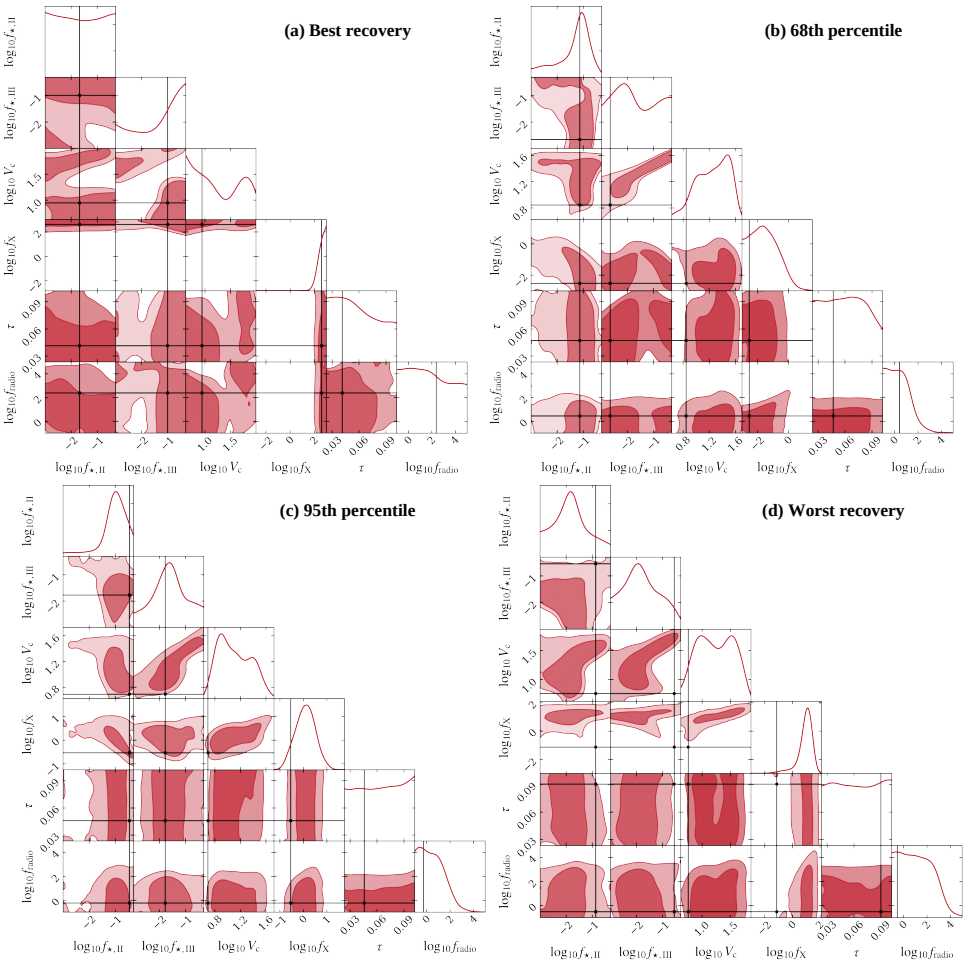


Figure 4: **Astrophysical parameter posteriors.** Marginal posteriors for the six signal parameters in the best, 50th-, 68th-, 95th-percentile, and worst cases ordered by minimum whole-profile credible level. Black lines and markers denote the injected parameters; dark and light red contours contain 68% and 95% posterior probability, respectively.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately state the paper's contributions and restrict the empirical claims to the observational configuration and uncertainty sources evaluated here. Broader observing windows and beam configurations have been evaluated in previous work and are discussed with the relevant citations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper provides a dedicated Limitations and Future Work section (Section 4), which discusses omitted observational and receiver-calibration effects, the increased posterior complexity expected from their inclusion, and the computational scaling limitations of the current sampling approach.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not claim new theorems or formal theoretical results. The underlying CEM and inhomogeneous radio-background simulators are established models described in the respective cited papers; the novel surrogate construction and analytical marginalisation used here are specified in full in the main text and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The main text and appendix specify the forward-model components, parameter priors, data-generation procedure, emulator training, beam-surrogate construction, marginalised likelihood, sampling configuration, and recovery metrics required to reproduce the principal results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The benchmark implementation and instructions are publicly available on GitHub. The synthetic benchmark datasets and inference products supporting the paper, including the NSS chains and production results, are publicly archived in Zenodo record 22046858. The original `21cmSPACE` training simulations and full CEM beam-generation and data-generation pipelines remain under private licences but are available upon reasonable request where permitted. These upstream assets are not required to rerun the released benchmark because the required emulator and response tensors are included.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main text and appendix report the data construction and splits, model architecture, optimiser, learning-rate schedule, training duration, parameter priors, beam-surrogate construction, and Nested Slice Sampling configuration used in the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Uncertainty is quantified directly by the Bayesian posterior distributions returned by Nested Slice Sampling. We report pointwise 95% posterior credible bands for each recovered signal, while variation across experiments is summarised by the RMSE distribution and its percentiles over 100 held-out signal realisations.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides dedicated timing benchmarks and reports the CPU and GPU architectures, available memory, per-evaluation and per-run runtimes, training time, and aggregate compute required for the simulation ensembles and 100 inference runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses simulated astrophysical, instrumental, and radiometer spectra and involves no personal data, human participants, surveillance, profiling, or consequential automated decision-making.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: By reducing simulation and inference costs, the framework broadens access to statistically rigorous uncertainty propagation, supports reproducible comparison of inference algorithms, and enables more reliable interpretation of global 21-cm observations. The work uses synthetic scientific data and is not expected to create material privacy, fairness, or dual-use risks.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The scientific simulator, synthetic spectra, and inference products contain no sensitive information and present no credible high-risk misuse pathway.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing software, simulation products, data, and models are correctly credited through the relevant citations and used with permission under their applicable terms. The benchmark code and documentation are released on GitHub under the BSD 3-Clause License, with the pretrained emulator weights provided under a separate 21cmSPACE Benchmark Model Use Licence. The benchmark datasets and inference products are archived on Zenodo under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The hybrid forward-model framework, synthetic benchmark datasets, and inference products introduced in this work are documented in the main text, appendix, and accompanying repository. The released materials describe their construction, parameterisation, provenance, structure, and limitations and are archived in the GitHub repository and Zenodo record cited above. No human-derived data are included.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No crowdsourcing, participant recruitment, or human-subject data collection was undertaken.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The study contains no human participants or identifiable human data and therefore does not require institutional human-subject review.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used only to refine text and assist with code debugging. They were not used in the simulator, model construction, inference methodology, data analysis, or experimental evaluation, and are not a component of the scientific method presented here.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.