Optimising Foreground Parametrisation for Global 21cm Cosmology with GPU-Accelerated Nested Sampling

Jacob L. Tutt[1,2★] Peter H. Sims[1,2] Dominic J. Anstey[1,2] Joe H. N. Pattison[1,2]
Eloy de Lera Acedo[1,2] and Samuel A. K. Leeney[1,2]
[1]Astrophysics Group, Cavendish Laboratory, J. J. Thomson Avenue, Cambridge, CB3 0HE, UK
[2]Kavli Institute for Cosmology, Madingley Road, Cambridge, CB3 0HA, UK

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

The global 21-cm signal from the Cosmic Dawn and Epoch of Reionisation provides an unparalleled probe of the astrophysics governing the early Universe, yet its detection is severely hindered by Galactic foregrounds that are many orders of magnitude brighter than the signal of interest. Chromatic distortions introduced by the antenna beam further complicate signal recovery, requiring highly accurate foreground modelling, rigorous Bayesian model comparison, and robust validation frameworks. In this work, we first demonstrate a substantial acceleration of Nested Sampling enabled by parallelisation on GPU architectures, achieving reductions in wall-time and computational cost of $O(10^2$–$10^3)$. Leveraging this increased computational capability, we introduce a novel observation-dependent sky-partitioning scheme that dynamically defines foreground regions using the antenna beam-convolved sky power for a given observing window. We show that this scheme improves modelling performance through three key avenues. First, by enforcing a strictly nested region hierarchy that enables clear identification of the Occam penalty in the Bayesian evidence, facilitating statistically principled optimisation of model complexity. Second, by enabling more accurate inference of spatially varying spectral indices, with posterior estimates consistently centred within true physical ranges and, thirdly, by enabling complex Galactic foregrounds to be modelled at the accuracy required for robust global 21-cm signal recovery using a significantly smaller parameter set.

Key words: methods: data analysis – methods: statistical - dark ages, reionization, first stars – cosmology: observations

1 INTRODUCTION

Driven by a wealth of high-resolution observations across the electromagnetic spectrum, the fields of astrophysics and cosmology have rapidly transformed into data-rich disciplines, allowing ever tighter constraints to be placed on the physical mechanisms and fundamental parameters governing the Universe's evolution. At high-redshift ($z \approx 1100$), observations of the Cosmic Microwave Background (CMB) (Smoot et al. 1992; Bennett et al. 2003; Fowler et al. 2010; Planck Collaboration et al. 2014, 2016, 2020) emitted during recombination provide a high-precision snapshot of the density anisotropies within the infant Universe. As these gravitational instabilities imprinted on the primordial blueprint collapsed (Bernardeau et al. 2002), they eventually resulted in the distribution of nearby galaxies within the cosmic web seen today. These low-redshift structures can be probed in similarly exquisite detail by large-scale spectroscopic surveys, such as the Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al. 2013) and the Dark Energy Spectroscopic Instrument (DESI; Adame et al. 2025).

Despite these two well-explored bookends, the vast majority of cosmic history remains unmapped, most notably the Cosmic Dark Ages (DA, $z \sim 1100 - 30$); the Cosmic Dawn (CD, $z \sim 30 - 20$) and the Epoch of Reionisation (EoR, $z \sim 20 - 6$). Towards this aim, the high resolution ($\lesssim 0.1$ arcseconds) of the recently launched James Webb Space Telescope (JWST; Gardner et al. 2006) is revolutionising our understanding deep within the EoR by directly imaging some of the earliest galaxies ($z \gtrsim 10$). By revealing an excess of bright ancient galaxies relative to prior predictions (Whitler et al. 2025), the mission is already prompting a re-evaluation of the theoretical models governing the Cosmic Dawn (see Li, Zhaozhou et al. 2024; Hutter, Anne et al. 2025; Kravtsov & Belokurov 2024). However, JWST's small survey volumes, coupled with the intrinsic faintness of the first luminous sources ($z \sim 20 - 30$), placing them beyond reach, mean it will struggle to stringently constrain the astrophysical parameters of these early epochs.

In comparison, the tracing of neutral hydrogen's redshifted 21-cm hyperfine transition from within the intergalactic medium (IGM) promises to provide statistically robust insights into key properties such as the initial mass function (Gessey-Jones et al. 2022), formation efficiency, and spectral emissivity (Schauer et al. 2019) of Population III stars as well as the nature of their associated X-ray binaries (Sartorio et al. 2023) (for comprehensive reviews, see Furlanetto et al. 2006; Pritchard & Loeb 2012; Barkana 2016; Mesinger 2019). Beyond these, the 21-cm signal has also been shown to encode exotic physics, including the contribution of primordial black holes (Mittal et al. 2022), interacting dark matter models (Barkana 2018), and superconducting cosmic strings (Brandenberger et al. 2019; Gessey-Jones et al. 2024) to a potential excess radio back-

★ E-mail: jlt67@cam.ac.uk

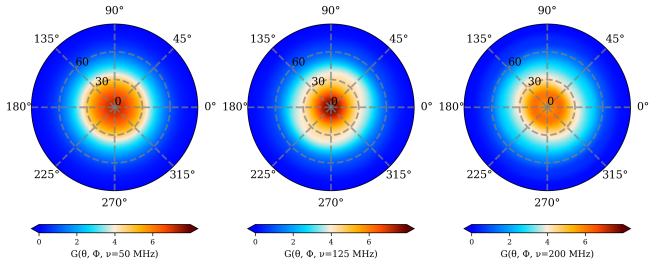

Figure 1. Beam pattern for a conical log spiral antenna, shown as polar projections of the antenna gain $G(\theta, \phi, \nu)$ in local altitude–azimuth coordinates (θ, ϕ) above the horizon ($\theta < 90°$). The three panels show $G(\theta, \phi, \nu)$ at ν = 50, 125 and 200 MHz to demonstrate the chromatic structure of the beam.

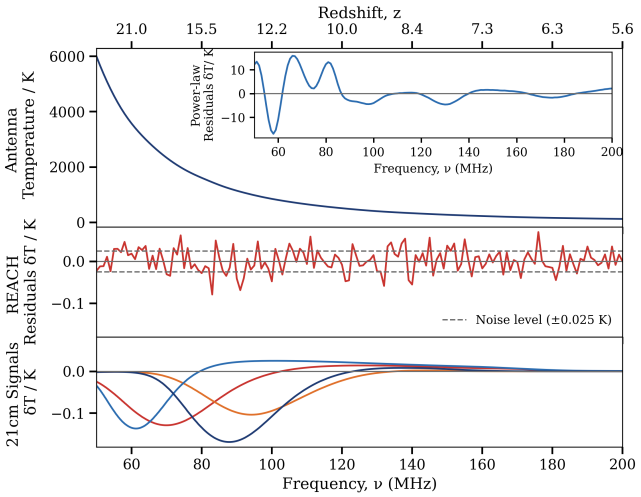

Figure 2. A demonstration of the chromatic structure introduced by the coupling of the Galactic foregrounds and the beam, and how they can be accounted for with the REACH data analysis pipeline. Top panel: A simulated 1 hour time-averaged observation $d(\nu)$ from the REACH telescope in the Karoo Desert, South Africa at 00:00 01-10-2019 with a conical log spiral antenna. The data has a mock 21-cm signal and 0.025 mK of Gaussian noise injected. The inset shows residuals beyond a fitted smooth power-law. Middle panel: Reduced residual structure after subtracting a Bayesian nested sampling fit model produced by the REACH pipeline using 16 parametrised regions. Bottom panel: Emulated mock global 21-cm signals from GlobalEMU Bevins et al. (2021), to demonstrate the success of beam-aware modelling suppressing residuals below the magnitude of expected global 21-cm signal.

ground. Current experimental approaches for 21-cm Cosmology can be broadly classified into two categories. Firstly, interferometric arrays such as HERA (DeBoer et al. 2017), LOFAR (van Haarlem, M. P. et al. 2013), the MWA (Tingay et al. 2013) and the upcoming SKA-Low (Mellema et al. 2015) which aim to measure spatial fluctuations in the 21-cm brightness temperature through its power spectrum and, ultimately, tomographic imaging. Secondly, as a complementary approach, there exists a wealth of Global 21-cm experiments focusing on the sky-averaged brightness including EDGES (Experiment to Detect the Global Epoch of Reionisation Signature Bowman et al. 2018), PRIZM (Probing Radio Intensity at High-Z from Marion Philip et al. 2018), SARAS (Shaped Antenna measurement of the background RAdio Spectrum Singh et al. 2018), MIST (Mapper of the IGM Spin Temperature Monsalve et al. 2024) and REACH (Radio Experiment for the Analysis of Cosmic Hydrogen de Lera Acedo et al. 2022). While the EDGES collaboration has claimed the first tentative detection of the global 21-cm signal, its large amplitude ($\approx$ 500 mK), low central frequency ($\approx$ 78 MHz), and flattened Gaussian profile have been interpreted as either evidence for physics beyond standard ΛCDM cosmology (Reis et al. 2021; Liu et al. 2019), or as the result of residual, unmodelled systematics in the data analysis (Hills et al. 2018; Singh & Subrahmanyan 2019; Sims & Pober 2019; Bevins et al. 2021). The latter, non-cosmological interpretation is further supported by the SARAS3's null detection, which rules out the EDGES absorption profile with 95.3 per cent confidence (Singh et al. 2022).

While detecting the global 21-cm signal faces a number of challenges, ranging from the mitigation of radio frequency interference (RFI; Fridman & Baan 2001) to distortions introduced by the ionosphere (Datta et al. 2016; Shen et al. 2021), the primary hurdle remains accurately accounting for the Galactic and extragalactic foregrounds emission. These contaminating foregrounds exceed the expected cosmological contribution by approximately three to four orders of magnitude across the relevant frequency range ($\simeq$ 50–200 MHz) (Shaver et al. 1999). Typically, experiments' foreground removal strategies exploit the comparatively spectrally smooth nature of synchrotron and free–free emission, allowing them to be modelled using power laws (Morales et al. 2006), log-polynomials (Harker et al. 2012), or derivative-constrained functions (Bevins et al. 2021). However, as demonstrated by Anstey et al. (2021), the spatial structure of Galactic foregrounds, when coupled with a chromatic antenna beam (see Figure 1), introduces frequency-dependent distortions that can become degenerate with the underlying cosmological signal (Figure 2).

To enable a statistically principled inference, the REACH collaboration introduced a Bayesian evidence–driven analysis framework (Anstey et al. 2021, 2023) that jointly models the convolution of sky realisations with the antenna beam alongside the 21-cm signal, thereby allowing parameter degeneracies and associated uncertainties to be accurately quantified (for a demonstration see Figure 2). As a consequence of this approach, the structure and spectral behavior of the low-frequency radio sky is simultaneously constrained (Carter et al. 2025), constituting an active area of research in its own right and a valuable resource for the broader community, independent of a confirmed detection of the global 21-cm signal (de Oliveira-Costa et al. 2008; Zheng et al. 2016; Dowell et al. 2017).

This paper focuses on optimising parameterised sky models in a physically motivated manner in order to improve the recovery of foreground spectral index parameters and thereby better mitigate foreground systematics to levels below those that impact cosmological signal recovery. This optimisation is benchmarked using a comprehensive Bayesian validation framework (Sims et al. 2025a), enabling a rigorous comparison of model performance. Specifically, we apply this framework to the REACH radiometer (Cumner et al. 2022), but the methodology is applicable to all physically motivated analyses of global 21-cm experiments.

The remainder of this paper is structured as follows. In section 2, we describe the Bayesian data analysis pipeline and its acceleration through the parallel processing capabilities of Graphics Processing Units (GPUs). section 3 outlines the limitations of existing sky parameterisation approaches and introduces the observationally dependent algorithm adopted in this work. The statistical validation framework and associated performance metrics used to draw comparisons between methods are presented in section 4. Finally, the results of applying this framework to simulated datasets are reported in section 5, and conclusions are drawn in section 6.

2 BAYESIAN DATA ANALYSIS PIPELINE

In this section, we present an overview of the methodology used to accurately simulate the antenna temperature for a given observation (subsection 2.1), together with the parameterised forward models (subsection 2.2) and Bayesian inference sampling algorithms (subsection 2.3, subsection 2.4) employed to efficiently solve the associated inverse problem. While the discussion below focuses on accounting for the diffuse foregrounds alongside the cosmological signal, the mathematical formalism required to incorporate additional physical effects, including RFI (Leeney et al. 2023; Anstey & Leeney 2024), the ionosphere (Shen et al. 2022), extragalactic point sources (Mittal et al. 2024) and environmental conditions (Pattison et al. 2023, 2025) is already established and can be incorporated in a modular manner. A schematic overview of the full analysis pipeline is shown in Figure 3.

2.1 Data Simulation

To fully capture the chromatic distortions introduced by diffuse foreground emissions into the observed data, simulations require full-resolution sky models that encode realistic spatial structure and frequency-dependent power distributions. These models will hereafter be referenced in relation to the local Alt–Az (θ, ϕ) coordinate frame of the antenna and thus a function of Coordinated Universal Time (UTC). Following Anstey et al. 2021, an observationally motivated realisation of such a model can be obtained using the 2008 Global Sky Model (GSM de Oliveira-Costa et al. 2008) evaluated at 408 MHz ($T_{408}(\theta, \phi, t)$) and 230 MHz ($T_{230}(\theta, \phi, t)$). By comparing the two frequencies, a spatially varying spectral index field $\beta(\theta, \phi, t)$ is derived as:

$$\beta(\theta, \phi, t) = \frac{\log[(T_{230}(\theta, \phi, t) - T_{\rm CMB})/(T_{408}(\theta, \phi, t) - T_{\rm CMB})]}{\log(230/408)}, \tag{1}$$

which is then used to extrapolate a reference sky at frequency ν_0 (taken here to be $T_{230}(\theta, \phi, t)$) to arbitrary observing frequencies, as shown in Equation 2. This procedure yields a continuous low-frequency sky model $T_{\rm sky}(\theta, \phi, \nu, t)$.

$$T_{\rm sky}(\theta, \phi, \nu, t) = [T_{230}(\theta, \phi, t) - T_{\rm CMB}] \left(\frac{\nu}{\nu_0}\right)^{-\beta(\theta,\phi,t)} + T_{\rm CMB}. \tag{2}$$

The simulated sky brightness temperature is then convolved with the directional gain of the antenna beam, $D(\theta, \phi, \nu)$, to produce the corresponding antenna temperature $T_{\rm data}(\nu, t)$. For demonstration purposes through this work, the antenna beam is modeled as a 6-m conical log-spiral antenna. To facilitate signal recovery tests, the mock data includes a realistic mock global 21-cm signal, $T_{21}(\nu)$. This is modeled as a Gaussian absorption profile, $\mathcal{M}_{21}(\theta_{21})$, with a central frequency (ν_{21}) of 85 MHz, a 15 MHz width (σ_{21}), and an amplitude of 0.155 K (A_{21}):

$$T_{21}(\nu) = A_{21}\ \exp\left[-\frac{(\nu - \nu_{21})^2}{2\sigma_{21}^2}\right]. \tag{3}$$

Finally, a noise realisation $\hat{\sigma}$ is added, which for the purposes of being comparable with prior work was assumed to be uncorrelated Gaussian noise with an amplitude of 25 mK:

Table 1. Parameters for the three observational windows used to benchmark foreground reconstruction and signal recovery, all starting at 00:00:00 on the respective date.

Reference	Date/ Duration	Configuration
Galaxy Down	01-10-2019 / 1 hr	Galactic pole at zenith.
Galaxy Up	01-07-2019 / 1 hr	Galactic center at zenith.
Galaxy 4hr	01-01-2019 / 4 hr	Integrated transit of the Galaxy.

$$T_{\rm data}(\nu, t) = \frac{1}{4\pi}\int_0^{4\pi} D(\theta, \phi, \nu)\, T_{\rm sky}(\theta, \phi, \nu, t)\, {\rm d}\Omega + T_{21}(\nu) + \hat{\sigma}. \tag{4}$$

Throughout this work, we benchmark our parameterised models against a range of simulated foreground complexities derived from three distinct observational windows with varying Galactic orientations, as detailed in Table 1 and visualised via the effective sky coverage shown in Figure 4. These shorter integration periods are intentionally selected to maximise the chromatic structure induced by the lack of sky rotation overhead; this is particularly pronounced in the 'Galaxy Up' case, where the power of the Galactic center further magnifies the amplitude of the distortions introduced. Consequently, the 'Galaxy Up' case represents an extreme scenario where 21-cm signal recovery would likely not be attempted in isolation on real data. However, it serves as a robust stress test of our dynamic models' ability to improve foreground recovery, and ensures that our validation metrics correctly flag reconstructions that are inadequate for the precision required for cosmological inference. Conversely, the 4-hour integration spans a broader range of Galactic positions, representing a more typical and viable observation target for signal inference. We note that while the equations presented throughout this section maintain the time-dependent t notation for mathematical generality, as time-resolved analysis is benchmarked in section A, the primary analysis in this work is performed on time-integrated data. This is simply achieved by collapsing the time domain in both the forward models and data simulations, creating a single integrated spectrum for each observational window.

2.2 Physically Motivated Foreground Model

While the simulation pipeline described in subsection 2.1 provides high-fidelity realisations, the calculated spatial distribution of both the base temperature $T_{408}(\theta, \phi, t)$ and the spectral index $\beta(\theta, \phi, t)$ is subject to observational uncertainties and thus likely offset from the true radio foregrounds. When aiming to perform signal inference from observational data, using these as fixed templates would introduce systematics into the modelled antenna temperature that prevent the recovery of the true global 21-cm signal. To mitigate this, the spectral indices and base-map amplitude must be simultaneously parameterised and incorporated within the foreground model to be jointly fit.

However, due to the nature of the one-dimensional data $T_{\rm data}(\nu)$ or two-dimensional data $T_{\rm data}(\nu, t)$ produced by a single radiometer, pixel-level parameterisation would be both highly degenerate and computationally prohibitive. Therefore, we adopt a regional parameter approach (Anstey et al. 2021, 2023; Pagano et al. 2023), which partitions the sky into N_β regions of uniform spectral index and N_α regions of uniform amplitude scaling. These are defined by the independent binary masks, $M_{\beta,j}$ and $M_{\alpha,i}$, which determine the membership of a pixel (θ, ϕ, t) to a specific region:

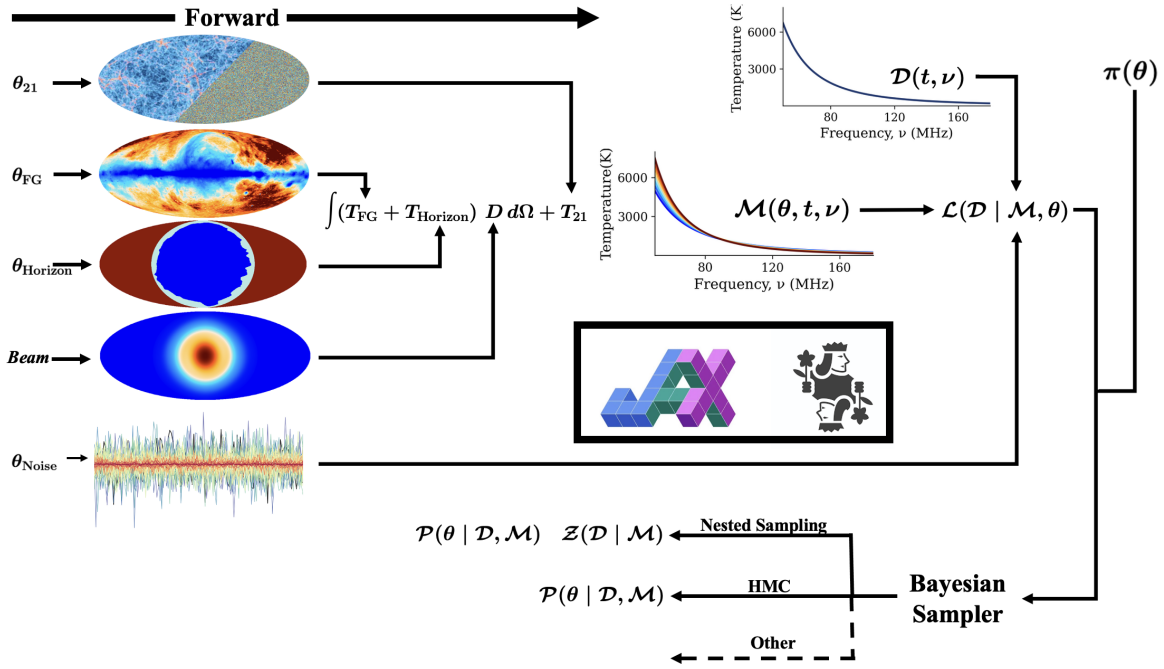

Figure 3. Schematic of the GPU-accelerated, differentiable REACH Bayesian analysis pipeline. The parameterised forward model combines a global 21-cm signal $\theta_{21\rm cm}$, diffuse foreground emission $\theta_{\rm FG}$ and horizon contamination $\theta_{\rm Horizon}$ with the antenna's beam to generate an antenna-temperature spectrum $\mathcal{M}(\theta, t, \nu)$. This model can be statistically compared to the observational data $\mathcal{D}(t, \nu)$ under a specific noise structure through a likelihood function $\mathcal{L}(\mathcal{D} \mid \mathcal{M}, \theta)$. The inference process is optimised through JAX's XLA compilation Bradbury et al. (2018), leveraging gradient-based `BlackJAX` samplers Cabezas et al. (2024) and nested-sampling variants Yallup et al. (2025a) for efficient Bayesian posterior $P(\theta \mid \mathcal{D}, \mathcal{M})$ and evidence $\mathcal{Z}(\mathcal{D} \mid \mathcal{M})$ evaluations.

$$T_{\rm sky}^{\rm model}(\theta,\phi,\nu,t) = \sum_{i=1}^{N_\alpha}\sum_{j=1}^{N_\beta} M_{\alpha,i}(\theta,\phi,t) M_{\beta,j}(\theta,\phi,t) \times \left[\alpha_i\,(T_{408}(\theta,\phi,t) - T_{\rm CMB})\right]\left(\frac{\nu}{\nu_0}\right)^{-\beta_j}, \tag{5}$$

where α_i is a multiplicative scale factor effectively acting as a localised gain correction to the 230 MHz template, and β_j is the fit spectral index for the $j^{\rm th}$ region.

To accelerate the forward model for a given observation window $\vec{t}$ and frequency band $\vec{\nu}$, the interaction between the beam, regional masks, and the reference base-map during the sky integration can be precomputed. This defines a chromatic response tensor, $\mathcal{K}_{i,j}(\nu,t)$, which allows the resultant antenna temperature, conditioned on any set of foreground parameters $\{\vec{\alpha},\vec{\beta}\}$, to be reduced to a simple series of matrix operations,

$$\mathcal{K}_{i,j}(\nu,t) = \frac{1}{4\pi}\int^{4\pi} M_{\alpha,i}(\theta,\phi,t) M_{\beta,j}(\theta,\phi,t) \times \left[T_{408}(\theta,\phi,t) - T_{\rm CMB}\right] D(\theta,\phi,\nu)\mathrm{d}\Omega\,, \tag{6}$$

and hence the antenna temperature is given by:

$$T_{\rm FG}(\nu,t) = \sum_{i=1}^{N_\alpha}\sum_{j=1}^{N_\beta} \alpha_i\,\mathcal{K}_{i,j}(\nu,t)\left(\frac{\nu}{\nu_0}\right)^{-\beta_j} + T_{\rm CMB}. \tag{7}$$

While previous implementations of this approach have relied on relatively simple partitioning schemes, this work focuses on optimising region definitions (section 3) to reconstruct continuous foregrounds with a minimal parameter set. This refinement is constrained by the condition that the models remain sufficiently expressive to ensure any residual systematics are statistically insignificant relative to the noise structure, a condition verified through a comprehensive validation suite (section 4). In the interest of clarity, this paper focuses specifically on the definition of spectral index masks ($M_{\beta,j}$); however, the algorithms introduced in section 3 are equally applicable to the amplitude scale factor masks ($M_{\alpha,i}$), the demonstration of which is left for future work.

2.3 Bayesian Inference

Given a forward model $\mathcal{M}$ describing the foregrounds and redshifted 21-cm signal through a set of parameters $\theta_{\mathcal{M}}$, we employ Bayesian inference to perform parameter estimation and model comparison for a given observational dataset $\mathcal{D}$. This is achieved by applying Bayes' theorem:

$$P(\theta_{\mathcal{M}} \mid \mathcal{D},\mathcal{M}) = \frac{\mathcal{L}(\mathcal{D} \mid \theta_{\mathcal{M}},\mathcal{M})\ \pi(\theta_{\mathcal{M}} \mid \mathcal{M})}{\mathcal{Z}(\mathcal{D} \mid \mathcal{M})}, \tag{8}$$

where $\pi(\theta_{\mathcal{M}} \mid \mathcal{M})$ denotes the prior probability distribution over the model parameters, encoding our initial state of knowledge (or lack thereof). The likelihood, $\mathcal{L}(\mathcal{D} \mid \theta_{\mathcal{M}},\mathcal{M})$, quantifies the probability of obtaining the observed data given a specific forward model, parameter set, and assumed noise structure. The resulting posterior, $P(\theta_{\mathcal{M}} \mid \mathcal{D},\mathcal{M})$, represents the updated probability distribution of parameters after incorporating the information contained in the data. Finally, $\mathcal{Z}(\mathcal{D} \mid \mathcal{M})$ is the Bayesian evidence, which measures the overall support for a model given the data and is defined as the likelihood marginalised over the full prior volume of parameters:

$$\mathcal{Z}(\mathcal{D} \mid \mathcal{M}) = \int \mathcal{L}(\mathcal{D} \mid \theta_{\mathcal{M}},\mathcal{M})\,\pi(\theta_{\mathcal{M}} \mid \mathcal{M})\,\mathrm{d}\theta_{\mathcal{M}}. \tag{9}$$

Table 2. Prior distributions for the global 21-cm signal, regional foregrounds, and instrumental noise parameters.

Parameter	Prior Dist.	Range	Units
Statistical Noise			
Noise Amplitude (σ_n)	Log-Uniform	$[10^{-4}, 10^{-1}]$	K
Regional Foregrounds			
Spectral Index (β_j)	Uniform	[2.458, 3.146]	–
Global 21-cm Signal			
Amplitude (A_{21})	Uniform	[0, 0.25]	K
Center Frequency (ν_{21})	Uniform	[50, 200]	MHz
Width (σ_{21})	Uniform	[10, 20]	MHz

2.3.1 Likelihood Function

In this work, we assume the noise $\hat{\sigma}$ is adequately described by a homoscedastic Gaussian distribution, however a host of more complex noise structures (Scheutwinkel et al. 2022a) including radiometric noise (Scheutwinkel et al. 2022b) have been previously explored in the context of 21-cm signal recovery. Given a dataset sampled across frequencies $\vec{\nu}$ and times $\vec{t}$, the log-likelihood ($\ln \mathcal{L}$) can thus be expressed as:

$$\ln \mathcal{L} = -\frac{1}{2} \sum_{k,l} \left[\ln(2\pi\sigma_n^2) + \frac{(T_{\text{data}}(\nu_k, t_l) - \mathcal{M}(\nu_k, t_l, \theta))^2}{\sigma_n^2} \right] \tag{10}$$

where $\mathcal{M}(\nu_k, t_l, \theta) = T_{\text{FG}}(\nu_k, t_l, \theta_{\text{FG}}) + T_{21}(\nu_k, \theta_{21})$ represents the combined foreground and signal forward model. Additionally, to account for the fact that in practical contexts, the exact noise properties are often not perfectly understood a priori, the Gaussian noise standard deviation, σ_n, is treated as a free parameter during inference.

2.3.2 Prior Distributions

The prior distributions utilised throughout our primary analysis (Table 2) were chosen to be sufficiently broad to encompass all physically plausible realisations of the low-frequency sky, therefore the spectral index priors were bounded by the extremes of the $\beta(\theta, \phi, t)$-map derived in Equation 1. For the global 21-cm signal, the prior ranges are informed by non-exotic astrophysical simulations (Cohen et al. 2017) from the semi-numerical code 21cmSPACE (Fialkov et al. 2013, 2014), reflecting the expected structure of signals.

2.3.3 Model Selection

Given the extreme sensitivity of global 21-cm signal recovery to mismodelling, it is critical to quantitatively demonstrate that any observational dataset $\mathcal{D}$ used for a claimed detection most strongly supports a signal-plus-foreground model ($\mathcal{M}_i$) against a foreground-only or alternative residual systematics model ($\mathcal{M}_j$). In order to perform this robust model comparison, evaluating the Bayesian evidence is essential as the relative probability of two competing models, i.e., $\mathcal{M}_i$ and $\mathcal{M}_j$, is determined by the ratio of their posterior odds, R_{ij}. By applying Bayes' theorem at the model level and assuming a non-informative (uniform) prior belief between them, $\pi(\mathcal{M}_1) = \pi(\mathcal{M}_2)$, this reduces to the ratio of their evidences, known as the Bayes Factor, B_{ij}:

Table 3. Interpretive mapping between the log Bayes factor $\ln(B_{ij})$ and qualitative levels of support for model $\mathcal{M}_i$ relative to $\mathcal{M}_j$.

$\ln(B_{ij})$	Odds in favour of $\mathcal{M}_i$	Preference for $\mathcal{M}_i$
$0 \leq \ln(B_{ij}) < 1$	1 – 3	Weak
$1 \leq \ln(B_{ij}) < 3$	3 – 20	Moderate
$3 \leq \ln(B_{ij}) < 5$	20 – 150	Strong
$\ln(B_{ij}) \geq 5$	> 150	Decisive

$$R_{ij} = \frac{P(\mathcal{M}_i \mid \mathcal{D})}{P(\mathcal{M}_j \mid \mathcal{D})} = \underbrace{\frac{\mathcal{Z}(\mathcal{D} \mid \mathcal{M}_i)}{\mathcal{Z}(\mathcal{D} \mid \mathcal{M}_j)}}_{B_{ij}} \underbrace{\frac{\pi(\mathcal{M}_i)}{\pi(\mathcal{M}_j)}}_{\text{prior odds}}. \tag{11}$$

To interpret the quantitative strength of preference implied by the Bayes factor B_{ij} (or equivalently R_{ij}), we adopt the qualitative classification scheme established by Kass & Raftery (1995), under which the degree of support for a given comparative model is categorised according to the ranges summarised in Table 3.

While gradient-based Markov Chain Monte Carlo (MCMC) methods, such as Hamiltonian Monte Carlo, excel at efficiently exploring the posterior topology to identify parameter correlations (Duane et al. 1987; Neal 1996; Hoffman & Gelman 2014), they do not natively provide a means to calculate the Bayesian evidence, $\mathcal{Z}$. To address this, we employ Nested Sampling (NS, Skilling 2006), which through evaluating the high-dimensional integral shown in Equation 9 simultaneously yields posterior samples and an accurate estimate of the evidence. However, as the dimensionality of the parameter space and the volume of observational data increases traditional CPU-based implementations of NS face significant scalability challenges. In this work alone, the systematic exploration and validation of various foreground partitioning schemes (see section 4) necessitates $O(10^3)$ independent nested sampling runs. Given this, reducing the overall inference cost, in terms of wall-time and computational resources is essential for feasible and reproducible analysis. To this end, we integrate the `BlackJAX` nested sampling framework (Yallup et al. 2025a; Cabezas et al. 2024), described in subsection 2.4, into the analysis pipeline.

2.4 GPU-Accelerated Nested Sampling

The parallel architecture of Graphics Processing Units (GPUs) enables substantial acceleration of nested sampling through two complementary mechanisms. The first is the hardware-level vectorisation of likelihood evaluations, detailed in subsubsection 2.4.1, while the second is an algorithmic reformulation of the nested sampling procedure, outlined in subsubsection 2.4.2. The discussion below focuses on an overview of the technical framework introduced by Yallup et al. (2025a), while benchmarking of the resulting performance gains is presented in section A. While not a primary focus of this work, it is worth noting that implementing the analysis pipeline within a `JAX`- and GPU-compatible framework enables automatic differentiation (Baydin et al. 2017), allowing gradients of the forward model and likelihood to be obtained at negligible additional computational cost. This capability naturally supports future extensions incorporating gradient-informed nested sampling schemes (Betancourt 2011; Lemos et al. 2024), as well as the integration of physically motivated forward models within Bayesian machine learning pipelines (see Saxena et al. 2024; Leeney et al. 2026).

2.4.1 Likelihood Parallelisation

As outlined above, following appropriate pre-computation, such as the construction of chromatic response tensors $\mathcal{K}_{i,j}(\nu, t)$, the likelihood reduces to a sequence of batched linear algebra operations. This structure maps naturally onto GPUs, which are explicitly designed to perform large-scale linear algebra workloads efficiently.

GPUs comprise thousands of lightweight cores optimised for high-throughput, Single Instruction–Multiple Data (SIMD) workloads, in contrast to CPUs, which prioritise low-latency, sequential performance (Owens et al. 2008). When coupled with `JAX`'s Accelerated Linear Algebra (XLA) compilation (Sabne 2020), they allow likelihood evaluations to be executed concurrently across many threads, yielding substantial reductions in wall-time. As a result, the effective scaling with increasing data volume or model dimensionality is strongly suppressed, approaching $O(1)$ behaviour, until limited by memory bandwidth.

2.4.2 Algorithmic Parallelisation

The second level of acceleration involves an adaptation of the Nested Sampling algorithm itself. Originally proposed by Skilling (2006), NS solves the evidence integral by mapping the high-dimensional parameter space Θ to a one-dimensional prior mass ξ, defined as the fractional volume of the prior where the likelihood exceeds a threshold $\mathcal{L}^*$:

$$\xi(\mathcal{L}^*) = \int_{\mathcal{L}(\theta)>\mathcal{L}^*} \pi(\theta)\, \mathrm{d}\theta. \tag{12}$$

Under this transformation, the evidence is simply reduced to a one-dimensional integral over prior mass. In conventional CPU implementations, this integral is evaluated numerically by evolving a population of n_{CPU} live points, with each iteration contracting the remaining prior mass through replacement of the point with the lowest likelihood. This process generates a discrete set of discarded (dead) points which allow the integral to be approximated via a weighted summation:

$$\mathcal{Z} = \int_0^1 \mathcal{L}(\xi)\, \mathrm{d}\xi \approx \sum_i (X_{i-1} - X_i)\mathcal{L}_i. \tag{13}$$

Here, $\mathcal{L}_i$ is the likelihood of the i^{th} discarded point, and the quadrature weight $(X_{i-1} - X_i)$ is determined by the stochastic contraction of the prior volume. For a population of n_{CPU} live points, the expected log-volume remaining after k iterations is given by:

$$\mathbb{E}\left[\log X_{\mathrm{CPU}}\right] = -\sum_{j=1}^{k} \frac{1}{n_{\mathrm{CPU}}} = -\frac{k}{n_{\mathrm{CPU}}}. \tag{14}$$

Following Yallup et al. (2025a), this inherently sequential process can be parallelised by simultaneously discarding the k lowest-likelihood points and launching independent slice-sampling chains in parallel across the GPU, each subject to the constraint $\mathcal{L} > \mathcal{L}_{\min,k}$, where $\mathcal{L}_{\min,k}$ is the maximum likelihood of the discarded batch. Accounting for the effective decrease in the number of live points throughout the batch, the expectation of the cumulative log-volume contraction is then:

$$\mathbb{E}\left[\log X_{\mathrm{GPU}}\right] = -\sum_{i=0}^{k-1} \frac{1}{n_{\mathrm{GPU}} - i} \approx \ln\left(\frac{n_{\mathrm{GPU}} - k}{n_{\mathrm{GPU}}}\right). \tag{15}$$

For further demonstrations of GPU-accelerated nested sampling applied to a broader range of cosmological and astrophysical inference problems, we refer the reader to Yallup et al. (2025b); Prathaban et al. (2025); Leeney (2025); Leeney et al. (2025) and Lovick et al. (2025).

2.4.3 Sampler Hyperparameters

The performance of the sampling algorithm is governed by a small number of key hyperparameters. These include the number of live points used throughout the run (`n_live`), influencing the resolution of the posterior and evidence estimation, the number of slice-sampling steps used to generate new live points (`num_inner_steps`), regulating the degree of correlation between samples and finally, the number of live points replaced simultaneously during each iteration (`num_delete`) controlling the level of algorithmic parallelism, trading computational efficiency against sampling accuracy. Throughout this work, we adopt `n_live` = 100 × `nDim`, `num_inner_steps` = 12 × `nDim`, and `num_delete` = 0.2 × `n_live`, where `nDim` denotes the number of free parameters in the given model configuration. These choices are informed by convergence and performance studies presented in section B.

3 REGION CONSTRUCTION

The accuracy with which the diffuse Galactic and extragalactic foregrounds can be approximated using the framework introduced in subsection 2.2 is intrinsically linked to both the number of regions adopted and the spatial logic used to define the corresponding masks.

If the partitioning is overly coarse, the assumption that extended regions of the sky can be modelled with a single spectral index or amplitude scale factor fails. Such under-parameterisation leaves systematic residuals in the modelled antenna temperature that may bias, or potentially mimic, the global 21-cm signal. Conversely, over-parameterisation through excessive subdivision incurs a prohibitive computational cost. Specifically, the number of required likelihood evaluations within slice-sampling-based NS algorithms (subsection 2.4) are highly sensitive to the dimensionality of the parameter space, in the worst case scaling as $O(D^3)$ (Handley et al. 2015), where D is the number of free parameters. Model compactness is therefore a critical consideration for practical implementation, especially given that robust analyses typically require large ensembles of inference runs for evidence-driven optimisation and validation.

This section first reviews the previously adopted sky-partitioning scheme and discusses its limitations (subsection 3.1), before introducing the methodology developed in this work to address these shortcomings (subsection 3.2 and subsubsection 3.3.1). A quantitative comparison between the two approaches is presented in section 5.

3.1 Traditional Partitioning and the Occam Penalty

Typically, regional partitioning has relied on static, observation-independent masks defined by dividing the spectral index map, $\beta(\theta, \phi, t)$ (Equation 1), into N_β uniform intervals of equal width (hereafter referred to as linear splitting).

One of the primary advantages of the Bayesian evidence $\mathcal{Z}$ is its intrinsic penalisation of additional model complexity that does not substantially improve the fit, a manifestation of Occam's razor (MacKay 1992). However, in the linear splitting scheme, increasing N_β to $N_\beta + 1$ does not simply add a new dimension to the existing posterior, it redefines all region boundaries across the sky and hence the entire parameter space.

Because the previous model configuration is not preserved as a nested subset of the new one, there is no guarantee that the $N_\beta + 1$ case

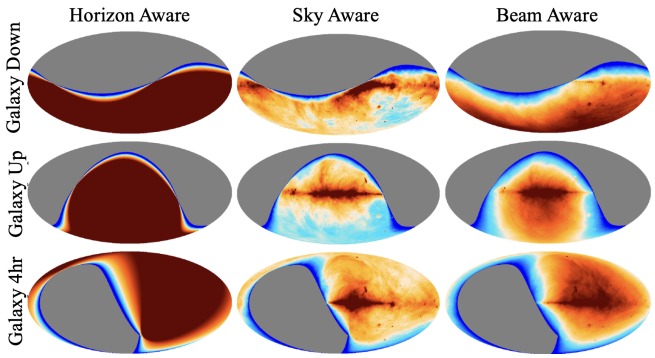

Figure 4. Time-integrated importance maps in Galactic coordinates used to construct sky regions. Each row corresponds to an observing window: Galaxy Down (1hr), Galaxy Up (1hr) and a 4hr Galaxy integration, and each column shows a different weighting: "horizon aware" (visibility-only), "sky aware" (sky-brightness-weighted), and "beam aware" (beam-convolved sky-brightness-weighted). The beam-aware maps concentrate weight where the instrument is both sensitive and the sky is bright, producing observation-dependent importance structures that motivate adaptive region definitions for foreground parametrisation.

will recover or exceed the maximum likelihood, $\mathcal{L}_{\max}$, of its predecessor. This inconsistency obscures the Occam's penalty, as fluctuations in evidence are driven simultaneously by changed spatial definitions and increased model flexibility. Furthermore, the loss of a monotonically increasing $\mathcal{L}_{\max}$, a robust indicator of algorithmic convergence as model complexity grows, removes a critical diagnostic for ensuring sufficient exploration of the increasingly high-dimensional likelihood surface.

Furthermore, because the sky brightness is spatially non-uniform and modulated by the antenna beam, regions defined under this method contribute unevenly to the observed antenna temperature $T_{\text{data}}(\nu)$. This lack of observational awareness leads to an inefficient allocation of degrees of freedom: parameters associated with regions of low beam-convolved sky brightness remain prior-dominated, inflating the dimensionality of the model space without significant improvement to the suppression of foreground systematics below that required for accurate signal recovery.

Tackling these limitations therefore requires addressing two primary challenges. First, considering the regional sky contributions to the observed antenna temperature, and second, ensuring that successive refinements maintain a nested partitioning structure. We propose a two-stage methodology to achieve this: the definition of an importance-weighted representation of the spectral index distribution (subsection 3.2), followed by the application of recursive algorithms to subdivide that distribution (subsection 3.3).

3.2 Spectral Index Importance Weighting

To better inform the construction of the foreground model for a given observation, we introduce an importance weighting map, $W(\theta, \phi, \nu, t)$. This quantifies the fractional relevance of each spatial coordinate to the total measured data, ensuring that the model's degrees of freedom are allocated where they are most justified.

In the following, we first outline a series of increasingly sophisticated weighting schemes, progressing from simple visibility constraints to complex instrument-aware sensitivity. We then describe how these high-dimensional maps are compressed into a one-dimensional cumulative distribution, $C(\beta', t)$, which serves as the foundation for defining region masks downstream.

3.2.1 Horizon-Aware Weighting

The most fundamental weighting scheme considers the visibility of the sky overhead given the constraints introduced by the local environment. By incorporating the static binary horizon mask $\mathcal{H}(\theta, \phi)$ (for full details see Pattison et al. 2023), this scheme identifies the subset of the spectral index distribution that is physically observable to the antenna at any given time. It weights the importance of each coordinate accordingly, ensuring the model is grounded by the observation's field of view across its Local Sidereal Time (LST) range:

$$W_{\text{Horizon}}(\theta, \phi) = \frac{\mathcal{H}(\theta, \phi)}{\int_{4\pi} \mathcal{H}(\theta', \phi')\, \mathrm{d}\Omega'} \,. \tag{16}$$

3.2.2 Sky-Brightness-Aware Weighting

The Horizon-Aware weighting can be further refined by incorporating the distribution of celestial power, $T_{\text{sky}}(\theta, \phi, \nu, t)$ (as defined in subsection 2.1). This scheme recognises that high-intensity regions, such as the Galactic center, exert a disproportionate influence on the total antenna temperature. By scaling the significance of each coordinate relative to its brightness, the resulting distribution ensures that the forward model's spatial resolution is concentrated on the regions that dominate the incident power:

$$W_{\text{Sky}}(\theta, \phi, \nu, t) = \frac{T_{\text{sky}}(\theta, \phi, \nu, t)\mathcal{H}(\theta, \phi)}{\int_{4\pi} T_{\text{sky}}(\theta', \phi', \nu, t)\mathcal{H}(\theta', \phi')\, \mathrm{d}\Omega'} \,. \tag{17}$$

3.2.3 Beam-Aware Weighting

While previous schemes assume uniform sensitivity across the visible sky, the final refinement incorporates the antenna beam, $D(\theta, \phi, \nu)$. Through calculating the convolution of the beam pattern with sky brightness across all LSTs, this weighting accounts for the spatial and frequency-dependent sensitivity of the instrument. Consequently, this provides the most accurate measure of the spatial foreground contribution to the observational data:

$$W_{\text{Beam}}(\theta, \phi, \nu, t) = \frac{T_{\text{sky}}(\theta, \phi, \nu, t) D(\theta, \phi, \nu) \mathcal{H}(\theta, \phi)}{\int_{4\pi} T_{\text{sky}}(\theta', \phi', \nu, t) D(\theta', \phi', \nu) \mathcal{H}(\theta', \phi')\, \mathrm{d}\Omega'} \,. \tag{18}$$

3.2.4 The Importance-Weighted Distribution

To transform the spatial importance maps into a form suitable for partitioning, they are compressed into a Cumulative Distribution Function (CDF) over the spectral index values. This distribution represents the total importance mass, establishing a principled basis for region definition where boundaries are informed by the relative observational contribution of different spectral index ranges. By integrating the weighting W across the sky, frequency, and time, we define $C(\beta', t)$ as:

$$C(\beta', t) = \int_{t,\, \nu,\, 4\pi} W(\theta, \phi, \nu, t)\, \mathbb{I}\left[\beta(\theta, \phi, t) \leq \beta'\right]\, \mathrm{d}\Omega\, \mathrm{d}\nu\, \mathrm{d}t \,, \tag{19}$$

where $\mathbb{I}$ denotes the indicator function. The necessity of the increasing

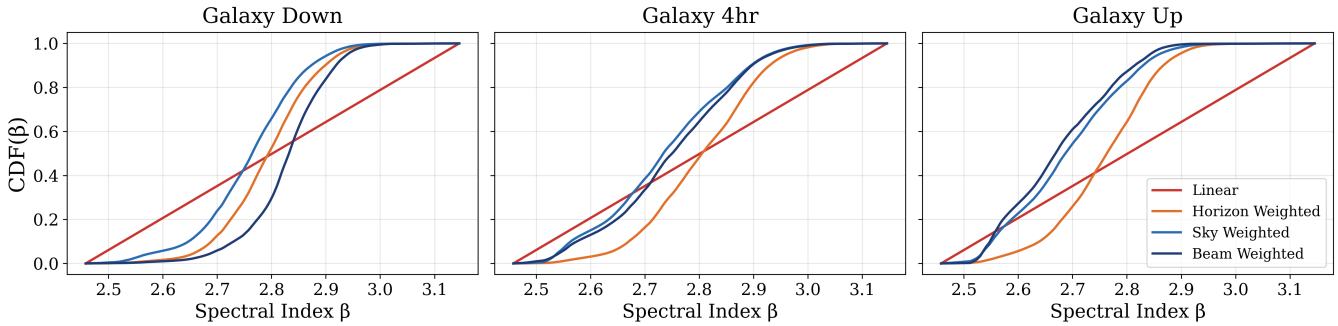

Figure 5. Spectral-index cumulative distribution functions (CDFs) constructed from different sky-weighting schemes. Each panel shows the empirical CDF of the diffuse-foreground spectral index β from a given observing window: Galaxy Down (1hr), Galaxy Up (1hr) and a 4hr Galaxy integration. Comparing an unweighted "linear" mapping (uniform in β) with horizon-, sky-, and beam-weighted CDFs constructed from the base spectral index map weighted by the corresponding time-integrated pixel importances shown in Figure 4.

levels of physical consideration can be demonstrated by the variations in the importance maps across the three observational windows, as illustrated in Figure 4 and their associated CDFs, detailed in Figure 5. We note that, while the CDFs constructed in this work are physically motivated, the region-construction framework is not restricted to this choice and may be generalised to alternative parametric forms, such as a beta distribution, enabling more flexible optimisation of the region definitions.

3.3 Algorithmic Partitioning Schemes

Given the importance-weighted CDF established in subsection 3.2, discrete sky regions are constructed by partitioning the resulting one-dimensional distribution optimally. We propose two distinct algorithmic frameworks, *Hierarchical Partitioning* (subsubsection 3.3.1) and *Recursive Partitioning* (subsubsection 3.3.2), differing in their strategy of allocating parameters across the spectral index domain. Crucially, both methodologies enforce a strictly nested model hierarchy enabling better informed model selection.

3.3.1 Hierarchical Partitioning

The Hierarchical scheme (Algorithm 1) first establishes a base partition of $N_{\text{base}} = 2^{\lfloor \log_2 N_{\text{total}} \rfloor}$ regions, each containing an equivalent fraction of the total importance mass ($\Delta C = 1/N_{\text{base}}$). For configurations where $N_{\text{total}} > N_{\text{base}}$, the algorithm resolves the remaining degrees of freedom by targeting regions with the greatest width in spectral index space, $\Delta\beta$ and subdividing them at their importance midpoints.

3.3.2 Recursive Partitioning

The Recursive scheme (Algorithm 2) uses iterative refinement. Initialised with a single region spanning the full spectral index range $[\beta_{\min}, \beta_{\max}]$, the algorithm recursively identifies the region encapsulating the highest importance density (ΔC) and bisects it at its midpoint in β-space.

3.3.3 Comparative Performance

While both algorithms seek to balance regions 'importance' and spectral index variance, they differ in their order of prioritisation.

Algorithm 1 Hierarchical Partitioning

Input: Total regions N_{total}, weighted CDF $C(\beta)$
Step 1: Base Partition Construction
$N_{\text{lower}} \leftarrow 2^{\lfloor \log_2(N_{\text{total}}) \rfloor}$ ▷ Largest power of $2 \leq N_{\text{total}}$
$\mathcal{R} \leftarrow$ Partition $C(\beta)$ into N_{lower} equal-mass regions
Step 2: Residual Resolution Enhancement
$N_{\text{rem}} \leftarrow N_{\text{total}} - N_{\text{lower}}$
while $N_{\text{rem}} > 0$ **do**
  Target: Identify $r^* = [\beta_a, \beta_b] \in \mathcal{R}$ with max width: $\Delta\beta_{r^*} = \beta_b - \beta_a$
  Split: Bisect r^* at its importance midpoint: $C_m = \frac{1}{2}(C(\beta_a) + C(\beta_b))$, $\beta_m = C^{-1}(C_m)$
  Update: $\mathcal{R} \leftarrow (\mathcal{R} \setminus \{r^*\}) \cup \{[\beta_a, \beta_m], [\beta_m, \beta_b]\}$
  $N_{\text{rem}} \leftarrow N_{\text{rem}} - 1$
end while
Return: $\mathcal{R}$

Algorithm 2 Recursive Partitioning

Input: Total regions N_{total}, weighted CDF $C(\beta)$
Step 1: Initialisation
$\mathcal{R} \leftarrow \{[\beta_{\min}, \beta_{\max}]\}$ ▷ Start with a single region
Step 2: Iterative Mass-Targeted Splitting
while $|\mathcal{R}| < N_{\text{total}}$ **do**
  Measure: For each $r \in \mathcal{R}$, calculate $m_r = \int_r \mathrm{d}C(\beta)$
  Target: Identify $r^* = [\beta_a, \beta_b] \in \mathcal{R}$ with max mass: $m_{r^*} = \max(\{m_r \mid r \in \mathcal{R}\})$
  Split: Bisect r^* at its physical midpoint: $\beta_m = \frac{1}{2}(\beta_a + \beta_b)$
  Update: $\mathcal{R} \leftarrow (\mathcal{R} \setminus \{r^*\}) \cup \{[\beta_a, \beta_m], [\beta_m, \beta_b]\}$
end while
Return: $\mathcal{R}$

Hierarchical Partitioning first ensures that each region has uniform importance mass and hence approximately equal contribution to the observed data, as evidenced in Figure 7. Conversely, Recursive Partitioning prioritises addressing the assumption that a given region can be defined by a single spectral index while balancing the requirement for higher parameter density where the foregrounds are most dominant.

In practice, while both algorithms showed advantages over the linear splitting method, the Recursive Partitioning framework offered a

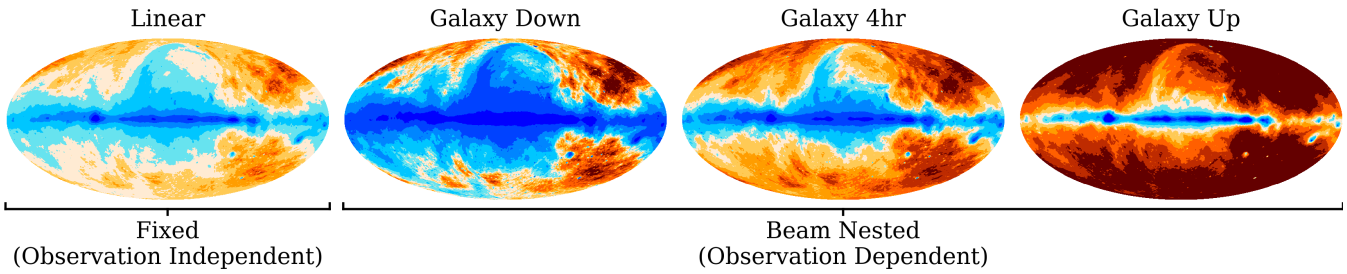

Figure 6. Comparison of discrete sky splitting masks $M_{\beta,j}$ for $N_{\text{reg}} = 11$ regions in Galactic coordinates. The columns compare an observation-independent linear splitting baseline with observation-dependent 'beam-weighted' recursive splitting schemes for three observing windows: Galaxy Down (1 hr), Galaxy Up (1 hr), and a 4-hr Galaxy integration. The adaptive scheme allows region boundaries to be redistributed to prioritise finer resolution in areas where the beam-convolved sky-brightness is highest.

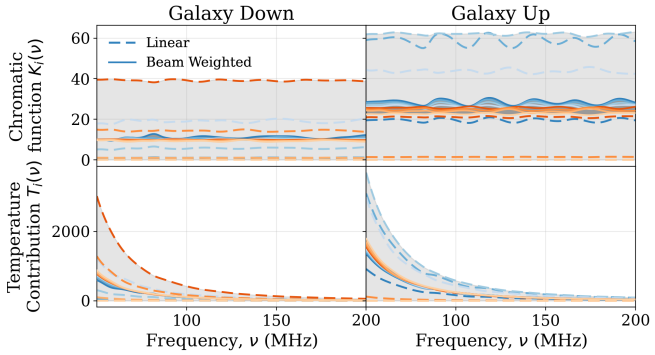

Figure 7. A comparison of the chromatic response tensors and spectral index scaled antenna temperature for 8 regions defined using linear versus beam-hierarchical sky splitting schemes. The columns show unique observing windows: Galaxy Down and Galaxy Up. Top row: chromaticity functions $K_i(\nu)$ for each region, shown for an observation-independent linear split (dashed) and an observation-dependent beam-hierarchical split (solid). Bottom row: corresponding region contributions $T_i(\nu)$ after applying assigned spectral-index scaling. The range of magnitudes/ contributions across all regions is demonstrated by the min/max shading for linear (light grey) and beam hierarchical (dark grey).

faster optimisation of the parameter space compared to the hierarchical alternative thus providing a superior ability to suppress systematic residuals while maintaining model compactness. Given this balance of accuracy and computational speed, it is the focus of the results presented hereafter. The impact of these observationally-dependent schemes on the resultant foreground masks, $M_{\beta,j}$, is illustrated in Figure 6. It provides a comparison between the 'Beam-Aware' Recursive scheme and the traditional linear baseline demonstrating the redistribution of regions based on the Galactic orientation.

4 STATISTICAL VALIDATION

While evaluating the significance of a potential detection via the Bayes factor is an essential step toward statistical rigor, the extreme sensitivity of signal recovery to foreground mismodelling means that, in isolation, it is insufficient to guarantee a physical detection. Given the aim of this work is to optimise the parametrisation of foreground models, it is essential to benchmark these refinements against a robust Bayesian validation framework. To address this, we adopt the methodology presented in Pattison et al. (2026, in prep), which describes the integration of the BaNTER validation framework (Sims et al. 2025a) into physically motivated Global-21cm analysis pipelines. In subsection 4.1, we discuss the specific failure modes of pure evidence-based comparisons, before defining the two key validation metrics used to identify and flag incorrect recoveries in subsection 4.2.

4.1 Failure Modes and Model Degeneracy

Although the detection Bayes factor, B_{det}, assesses whether the increased flexibility of a joint foreground-plus-signal model, $\mathcal{M}_{\text{FG+21}}$, yields a better statistical fit than a foreground-only model, $\mathcal{M}_{\text{FG}}$, it represents a blind comparison and is agnostic to the physical origin of that improvement. Due to inherent model-level degeneracies (Sims et al. 2025b) between the chromatic structure introduced by diffuse foregrounds and the global 21-cm signal, this metric cannot distinguish between cases where the signal recovery is reliable, or whether the 21-cm signal model is compensating for inaccuracies within the foreground model and hence the recovery is biased.

As an illustration of the range of possible inference outcomes, Figure 8 presents four representative cases of signal recovery in which the inclusion of a 21-cm signal model is decisively favoured according to the criteria in Table 3. Despite this strong statistical preference, only one case yields an accurate recovery of the injected signal parameters. Throughout this work, we quantify the accuracy of signal recovery using an uncertainty-aware metric defined in Equation 20, hereafter referred to as the Z-score, and classify recoveries with $Z < 1$ as accurate,

$$Z = \frac{1}{N_{\text{dim}}} \sqrt{\sum_{i=1}^{N_{\text{dim}}} \frac{(\mu_i - \theta_{i,\text{true}})^2}{\hat{\sigma}_i^2}}, \tag{20}$$

where μ_i is the posterior mean, $\theta_{i,\text{true}}$ is the true injected value, and $\hat{\sigma}_i$ is the posterior standard deviation for each of the N_{dim} signal parameters.

4.2 Validation Checks

To guard against the failure modes described above, we employ a two-stage validation strategy that jointly assesses signal–systematic

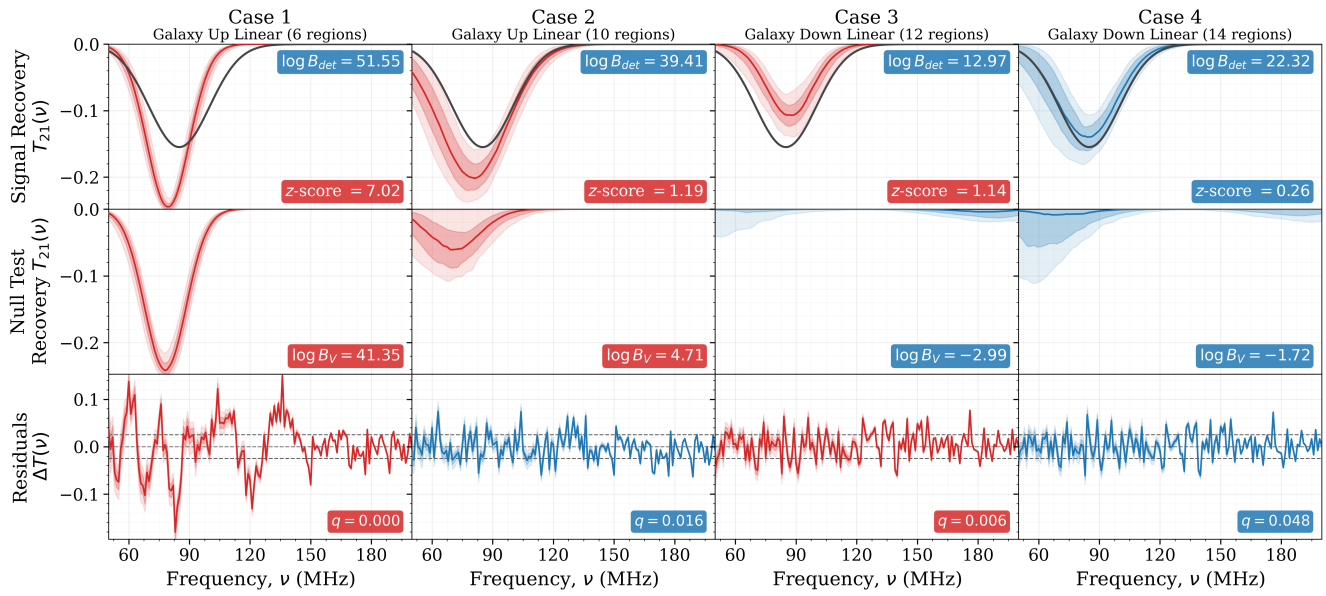

Figure 8. Illustration of the validation framework applied to representative signal-recovery scenarios. Each column corresponds to a distinct inference outcome, all of which exhibit a decisive statistical preference for the inclusion of a 21-cm signal model. Rows show (top to bottom) the recovered signal posterior for signal-injected data, the corresponding posterior obtained from the null test (no signal injected), and the residuals of the overall best-fitting model. Shaded bands indicate the posterior mean and 1σ and 2σ credible intervals. Blue and red curves denote configurations that pass or fail the respective validation criteria. Boxed annotations report the detection Bayes factor $\log B_{\rm det}$, the signal-recovery Z-score, and the null-test evidence ratio $\ln \mathcal{B}_{\rm V}$, highlighting cases in which statistically favoured detections nonetheless correspond to biased or unreliable signal recovery.

degeneracies through a null test (subsubsection 4.2.1) and evaluates the overall statistical consistency of a given fit through analysing the remaining residual structure (subsubsection 4.2.2).

4.2.1 *Null Tests*

The null test probes the susceptibility of a foreground model to spurious signal detection. We apply this test for a given observational period and parameterised foreground model by fitting the associated joint model, $\mathcal{M}_{\rm FG+21}$, to a validation dataset $D_{\rm V}$ that is intentionally simulated without an injected 21-cm signal. The resulting fit is then compared to a foreground-only model, $\mathcal{M}_{\rm FG}$, via the null-test evidence ratio:

$$\ln \mathcal{B}_V = \ln\left(\frac{\mathcal{Z}^{\rm V}_{\rm FG+21}}{\mathcal{Z}^{\rm V}_{\rm FG}}\right), \tag{21}$$

$\mathcal{Z}^{\rm V}_{\rm FG+21}$ and $\mathcal{Z}^{\rm V}_{\rm FG}$ denote the Bayesian evidences of the respective models. Due to the lack of signal within the validation data, any Bayes ratio favoring the composite model indicates that the signal component is compensating for residual foreground structure and therefore identifies foreground models that are insufficiently expressive and pose a high risk of biased detections when applied to observational data. Any configuration with $\ln \mathcal{B}_{\rm V} \geq 0$ is thus flagged accordingly.

4.2.2 *Residual Structure*

While the null test identifies problematic model configurations prior to their application to observational data, it does not assess whether an individual fit leaves residuals that are statistically consistent with instrumental noise. To do this, we compare the median a posteriori likelihood of a given posterior distribution, $\overline{\mathcal{L}_i}$, to the likelihood distribution expected if the data were described perfectly by the model up to random noise fluctuations. This reference distribution, denoted $\mathcal{L}_{\rm noise}$, is constructed by evaluating the likelihood using multiple realisations of the assumed noise model, taken in this work to be uncorrelated Gaussian noise with an amplitude of 25 mK.

The comparison is quantified by computing the fraction of the $\mathcal{L}_{\rm noise}$ distribution that yields likelihood values less than or equal to $\overline{\mathcal{L}_i}$:

$$q_i = \mathbb{P}\left(\mathcal{L}_{\rm noise} \leq \overline{\mathcal{L}_i}\right). \tag{22}$$

This represents the probability that a noise-only realisation produces residuals that are at least as well fit as those obtained from the model. Larger values of q_i therefore indicate that the residuals are statistically indistinguishable from noise, while smaller values signal the presence of coherent residual structure not captured by the model. We classify a fit, and hence the corresponding model, as statistically consistent if its median posterior likelihood lies within the upper $q_{\rm threshold}$ quantile of the ideal noise distribution, corresponding to $q_i \geq q_{\rm threshold}$. Throughout this work we adopt $q_{\rm threshold} = 0.99$. Fits failing this criterion are flagged as containing residual systematic structure, indicative of foreground mismodelling or signal–systematic degeneracy.

The requirement for both validation metrics to be applied in parallel is illustrated in Figure 8. While all cases shown strongly support the inclusion of a signal model (with $\ln B_{\rm det} \gg 0$), validation indicates that only Case 4 satisfies both criteria and can therefore be robustly trusted, corresponding to the sole accurate signal recovery. In particular, Case 2 demonstrates a scenario in which bias in the recovered 21-cm signal is sufficient to suppress foreground residuals such that they appear noise-like however this failure mode is still successfully identified by the null test. Conversely, Case 3 illustrates

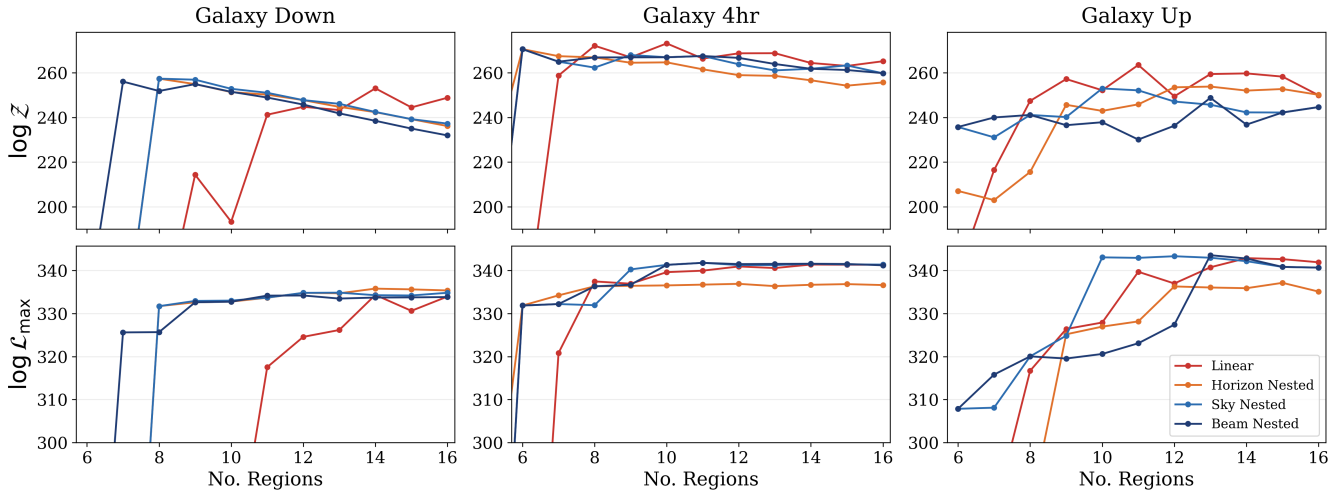

Figure 9. Bayesian evidence, $\log \mathcal{Z}$ (top row) and maximum log-likelihood, $\log \mathcal{L}_{\max}$ (bottom row) given by nested sampling versus the number of modeled foreground regions. Columns show Galaxy Down (1 hr), Galaxy Up (1 hr), and a 4-hr Galaxy integration, while each panel compares the four splitting strategies: linear, horizon-nested, sky-nested, and beam-nested.

a situation in which the null test is passed, yet residual systematics remain insufficiently Gaussian ($q < q_{\text{threshold}}$), rendering the inferred signal recovery unreliable.

5 RESULTS

In this section, we present the comparative results of a suite of nested-sampling fits across all three observational windows, explicitly benchmarking all importance-weighted partitioning schemes against the original observation-independent linear splitting. In response to the limitations of prior region definitions discussed in section 3, this section is structured as follows. In subsection 5.1, we examine how nested configurations improve Bayesian model comparison. We then demonstrate the degree to which the introduction of importance-aware region definitions better constrain foreground parameters in subsection 5.2. Finally, in subsection 5.3, we assess whether these improvements propagate to enhanced suppression of chromatic systematics, resulting in more reliable signal recovery.

5.1 Observing the Occam Penalty

As discussed in subsection 3.1, traditional linear splitting schemes do not define a consistent model hierarchy, as the reshuffling of region boundaries with increasing N_{reg} redefines the parameter space rather than expanding it. Figure 9 illustrates the consequences of this behaviour for both the maximum log-likelihood, $\log \mathcal{L}_{\max}$, and the Bayesian evidence, $\log \mathcal{Z}$, alongside the comparative advantages of the nested schemes introduced in this work.

First, the evolution of $\log \mathcal{L}_{\max}$ under the linear splitting scheme is highly non-monotonic, demonstrating that in the absence of a nested construction, increasing model dimensionality does not guarantee an improved description of the data. In particular, successive increases in N_{reg} within the linear scheme can introduce region definitions that are less optimal than those of lower-dimensional models. In contrast, all nested schemes exhibit a monotonic increase in $\log \mathcal{L}_{\max}$, up to small fluctuations attributable to algorithmic convergence. This behaviour guarantees that the introduction of additional parameters can only maintain or improve the quality of the fit.

Second, when considering the global maximum likelihood attained across the full range of region counts explored ($N_{\text{reg}} = 6$–16), the physically motivated sky- and beam-weighted schemes consistently outperform the linear baseline. This demonstrates that concentrating model flexibility in observationally significant regions of the sky yields foreground models that are more representative of the underlying structure. Moreover, these improvements are achieved more efficiently, supporting the Recursive partitioning strategy introduced in subsection 3.3. In particular, the importance-aware schemes reach a plateau in $\log \mathcal{L}_{\max}$ at substantially lower values of N_{reg} than the linear scheme requires to attain a comparable quality of fit.

This effect is most pronounced in the Galaxy Down observing window, where the likelihood saturates at $N_{\text{reg}} \simeq$ 8-9, while the linear scheme requires 14 or 16 regions to reach a similar level of performance. Although this trend is present across all observational windows, it is least pronounced for the Galaxy Up case, consistent with the strong Galactic emission and short integration times that necessitate higher model dimensionality for adequate resolution.

Finally, the nested constructions enable a much clearer identification of the Occam penalty through the Bayesian evidence. Because $\log \mathcal{Z}$ balances improvements in fit quality against the expansion of the prior volume, the saturation of $\log \mathcal{L}_{\max}$ in the Galaxy Down case is accompanied by a monotonic decline in $\log \mathcal{Z}$ beyond $N_{\text{reg}} \simeq$ 8-9. Under the linear splitting scheme, by contrast, the evidence exhibits no coherent trend, severely limiting its utility for principled model selection. While the evidence peak is less sharply defined for the Galaxy Up window, reflecting continued competition between improving fit quality and increasing prior volume, the nested schemes nonetheless exhibit the expected behaviour once their respective likelihood maxima are reached.

It is important to note, however, that the maximisation of the Bayesian evidence alone does not guarantee sufficient foreground recovery for robust 21-cm signal inference. As discussed previously, even small residuals bias the recovered global 21-cm signal, well below the scale at which they significantly impact $\log \mathcal{Z}$, motivating the validation framework introduced in section 4.

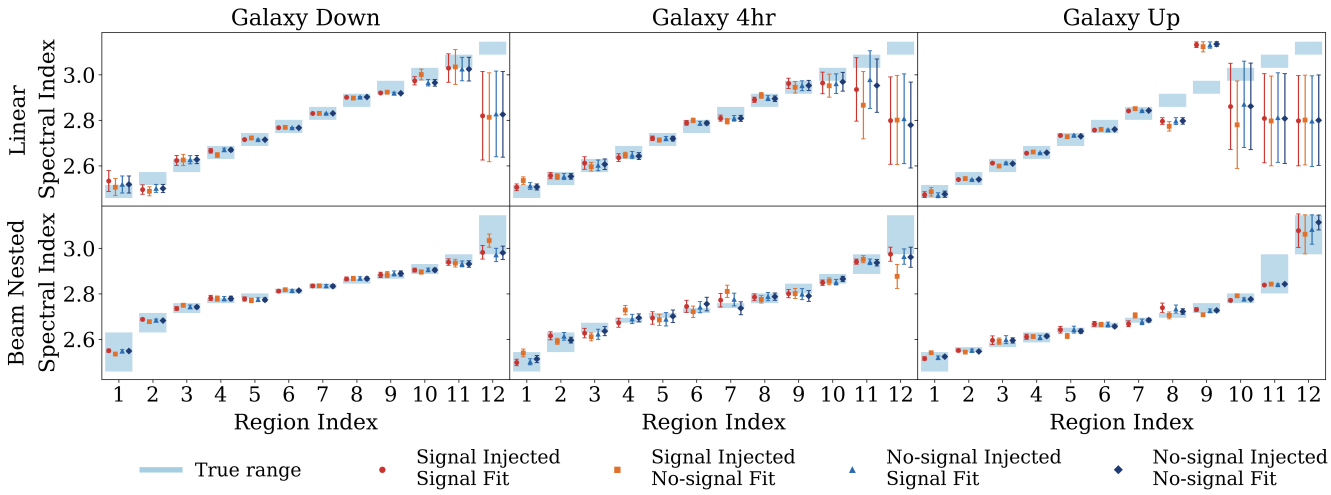

Figure 10. Comparison of foreground spectral index recovery using linear vs. beam-nested sky parameterisation schemes ($N_{\rm reg} = 12$). Columns represent different observing windows (Galaxy Down, Galaxy Up, and a 4 hr Integration), while rows compare the linear splitting method (top) with the beam-nested splitting method (bottom). In each panel, the shaded bands indicate the ground-truth range of spectral indices (β) present within that specific region on the high-resolution ($N_{\rm side} = 512$) base map used for data generation. The overlaid points and $1\,\sigma$ error bars show recovered posterior means for the four generation/fit combinations (Signal Injected/Signal Fit, Signal Injected/No-signal Fit, No-signal Injected/Signal Fit, No-signal Injected/No-signal Fit).

5.2 Foreground Recovery

Beyond the requirements of 21-cm signal extraction, the ability to update our prior state of knowledge regarding low-frequency radio maps and extrapolate them across a continuous frequency band is a valuable result in its own right. Figure 10 benchmarks the posterior spectral index recovery against the 'ground truth' ranges present in the full-resolution ($N_{\rm side=512}$) base map used in simulations for both the traditional linear splitting and the beam-nested scheme ($N_{\rm reg=12}$).

This comparison provides a clear visualisation of the dynamic range, $\Delta\beta$, captured by each mask under varying observational conditions. In the Galaxy Up case, for instance, the adaptive masks naturally cluster around lower spectral index values, correctly prioritising the flatter-spectrum emission characteristic of the Galactic plane. Critically, we find that the recovery of the spectral index values is far more robust under the new scheme. The original linear splitting exhibits two primary shortfalls: regions associated with low beam-convolved brightness remain prior-dominated (indicated by large uncertainties and poor centering), while others are recovered with high confidence but significant bias. The Galaxy Up case being particularly inaccurate example. In contrast, the beam-nested scheme consistently centres the posterior means within the true physical range across all regions.

Furthermore, we demonstrate the robustness of this foreground recovery across various permutations of signal injection and model fitting. As expected, the recovered foreground parameters remain largely independent of the underlying 21-cm cosmology, as the signal's amplitude is orders of magnitude below the chromatic distortions. The only marginal exception is the 4-hr Galaxy integration, where the longer integration period reduces the effects of beam-sky coupling such that the presence of a 21-cm signal without modeling leads to very slight offsets in the recovered spectral indices (see Signal Injected/ No Signal Fit). While Figure 10 focuses on the 12-region case, these performance gains are consistent across all investigated values of $N_{\rm reg}$.

5.3 Signal Recovery

Finally, we address the primary objective of this work, the capacity of the proposed modelling frameworks to pass the validation metrics and accurately recover the underlying 21-cm signal. Figure 11 summarises the results of the complete suite of nested sampling runs, with each cell reporting the detection Bayes factor ($\ln B_{\rm det}$), the null-test evidence ratio ($\ln \mathcal{B}_V$), and the overall signal recovery accuracy (Z-score). Each entry thus encompasses the outcome of four independent inference runs.

As a preliminary observation, these results support the robust success of the validation framework. In all investigated cases, the framework successfully identifies and flags inaccurate signal recoveries, a feature that is particularly relevant in scenarios exhibiting high Bayes factors that favor a detection. Here, the framework exposes the potential pitfalls of relying on $\ln B_{\rm det}$ in isolation, where decisive statistical support might otherwise lead to the acceptance of a biased parameter estimate.

In analysing the signal recovery, we examine the three observational windows independently. For the Galaxy Down case, the importance-weighted schemes enable accurate, validated signal recovery ($Z < 1$) using just 8 regions (for Sky and Horizon-Nested) or 9 regions (for Beam-Nested). In contrast, the original observation-independent linear splitting requires 14 regions to reach a validated recovery. A similar trend is observed for the 4-hour Galaxy integration, however, as this is intrinsically a less complex observation for signal detection, the required region counts are lower across all methodologies. While all importance-aware dynamic schemes achieve validated recovery with 7 regions, the linear splitting scheme requires 8.

The Galaxy Up case, included throughout this work as a high-stress test, presents more nuanced behavior. Interestingly, the linear splitting scheme shows marginal improvements in recovery at certain dimensionalities compared to the importance-weighted models, despite the latter's superior foreground parameter constraints discussed in subsection 5.2. Upon investigation, we find that while the

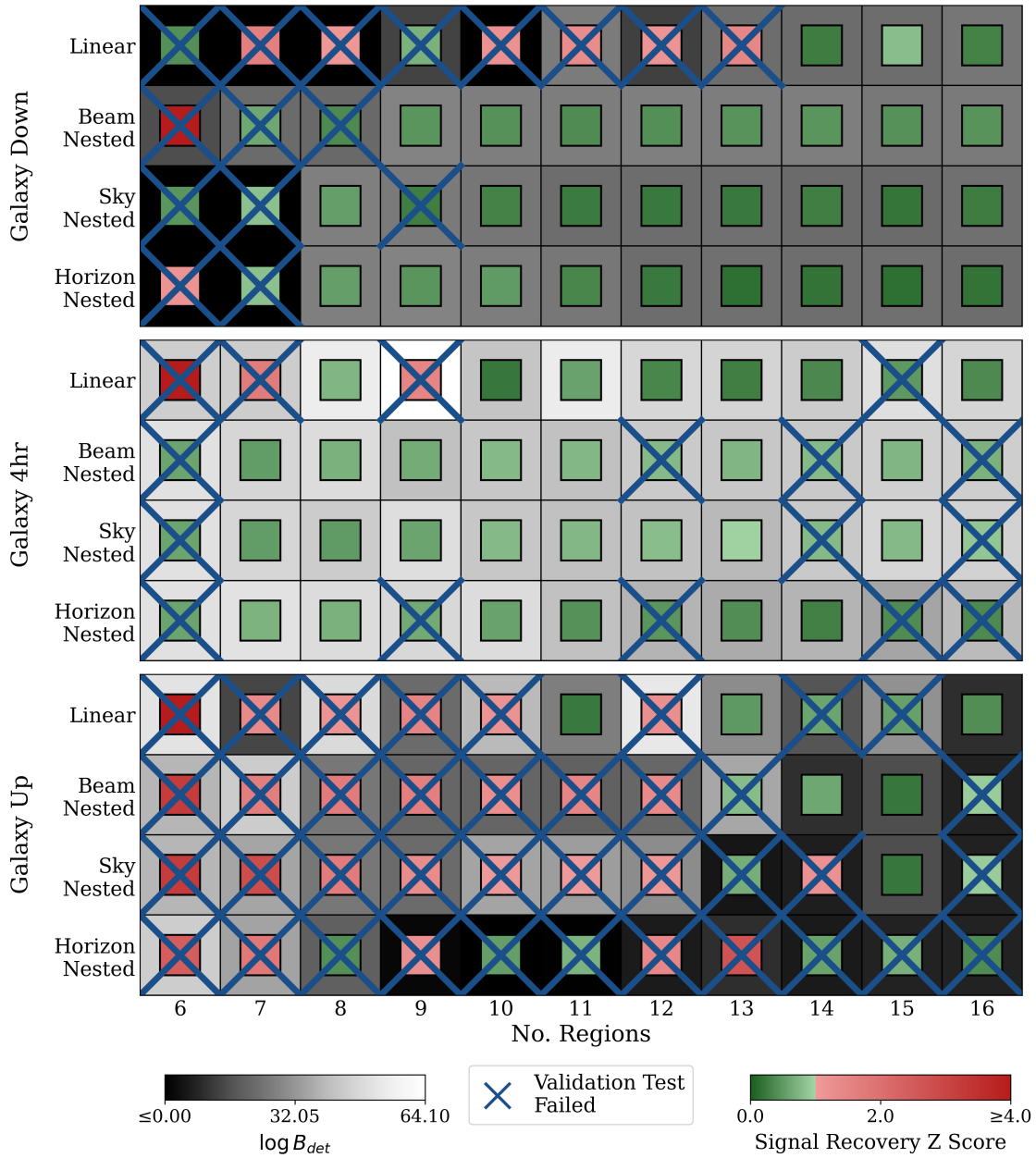

Figure 11. Summary of signal recovery and Bayesian model selection across sky-splitting methods and region counts. Each panel corresponds to a unique observing window, with columns indicating the number of regions ($N_{\rm reg}$ = 6–16) and rows showing the different sky-splitting schemes. The outer shaded square encodes the Bayesian evidence in favour of a signal detection, quantified by the log Bayes factor $\log B_{\rm det}$, with lighter colours indicating stronger support for the signal model. The inner square displays the statistical consistency of the recovered posterior with the injected signal via the Z-score, shown using a green–to–red colormap. Validation failures are indicated by a single blue cross, drawn when either the null test (signal preferred in no-signal injections) or the residual consistency test fails, signalling potential residual systematics degenerate with the global 21-cm signal.

signal recovery Z-scores are accurate, the flagging of the importance-weighted nested models is driven by the null test ($\ln \mathcal{B}_V \geq 0$) rather than the residual structure test. This level of signal-foreground degeneracy is a fundamental challenge that cannot be easily predicted a priori or accounted for within the physically motivated CDF construction and subsequent splitting scheme.

However, an intriguing avenue for future work would be to use the null-test Bayes factor itself as an optimisation criterion for region definition. Such an approach would require a parameterised, but not explicitly information-aware, CDF construction. Given exploring $O(10^2)$ alternative CDF configurations would be computationally prohibitive even with the GPU-accelerated nested-sampling pipeline presented in this work, such an approach would likely require simulation-based methodologies capable of providing rapid Bayesian model comparison forecasts. Recent developments in evidence networks (Jeffrey & Wandelt 2024; Gessey-Jones & Handley 2024) and conditional Bayesian Neural Ratio Estimation (cBNRE; Leeney et al. 2026) offer a promising route toward enabling such evidence-driven model optimisation. A detailed investigation of these approaches is left to future work.

6 CONCLUSIONS

In this work, we have presented a significant advancement in the computational and methodological framework for physically-motivated global 21-cm signal analysis. By leveraging GPU architectures together with modern compiler-based optimisation, we achieved a substantial acceleration of Nested Sampling Inference, reducing computational wall-time by factors of $O(10^2$–$10^3)$. This computational efficiency enabled the development and rigorous validation of a novel, observation-dependent sky-partitioning scheme designed to address the challenges of chromatic beam distortions and Galactic foreground contamination. Our results demonstrate that this dynamic partitioning improves foreground modelling through three primary avenues:

- **Principled Model Selection:** The enforcement of a strictly nested region hierarchy allows for the clear identification of the saturation of the maximum log-likelihood, $\ln \mathcal{L}_{\max}$ and consequently the Occam penalty within the Bayesian evidence, $\ln \mathcal{Z}$. This facilitates a statistically robust optimisation of model complexity, ensuring that the number of foreground regions is sufficient to accurately reconstruct the sky to the required precision without unnecessary over-parameterisation.
- **Improved Foreground Reconstruction:** The scheme yields more accurate recovery of spatially varying spectral indices. The resulting posterior distributions are consistently centred within true physical ranges, even in challenging observing windows such as when there is maximal coupling between the chromatic beam and high-intensity emission from the Galactic plane.
- **Efficient Signal Recovery:** Complex Galactic foregrounds can now be modelled at the precision required for robust global 21-cm signal recovery using a significantly smaller parameter set. This reduction in dimensionality, combined with our GPU-accelerated pipeline, makes high-fidelity Bayesian inference far less computationally expensive for large-scale experimental datasets.

While this study focused on spectral index masks within the REACH framework, the underlying algorithm is modular and adaptable. It can be readily applied to amplitude-based scale factor maps and is flexible enough to accommodate the diverse beam patterns and instrument responses of different global 21-cm experiments. Future work will explore the integration of this differentiable pipeline into advanced Bayesian and machine learning frameworks, providing a scalable path toward a confirmed detection of the 21-cm signal from the Cosmic Dawn and the Epoch of Reionisation.

ACKNOWLEDGEMENTS

The authors thank Will Handley for his contributions to the REACH pipeline as well as David Yallup for the development of BlackJax's Nested Sampling framework.

JT is supported by the Harding Distinguished Postgraduate Scholars Programme (HDPSP) and the Science and Technology Facilities Council (STFC) DTP Studentship.

We would also like to thank the Kavli Foundation for their support of REACH.

DATA AVAILABILITY

The data that supports the findings of this study are available from the first author upon reasonable request.

REFERENCES

Adame A., et al., 2025, Journal of Cosmology and Astroparticle Physics, 2025, 021
Anstey D., Leeney S. A. K., 2024, RAS Techniques and Instruments, 3, 372
Anstey D., de Lera Acedo E., Handley W., 2021, Monthly Notices of the Royal Astronomical Society, 506, 2041
Anstey D., de Lera Acedo E., Handley W., 2023, Monthly Notices of the Royal Astronomical Society, 520, 850
Barkana R., 2016, Phys. Rep., 645, 1
Barkana R., 2018, Nature, 555, 71
Baydin A. G., Pearlmutter B. A., Radul A. A., Siskind J. M., 2017, J. Mach. Learn. Res., 18, 5595–5637
Bennett C. L., et al., 2003, The Astrophysical Journal Supplement Series, 148, 97
Bernardeau F., Colombi S., Gaztañaga E., Scoccimarro R., 2002, Phys. Rep., 367, 1
Betancourt M. J., 2011, AIP Conf. Proc., 1305, 165
Bevins H. T. J., Handley W. J., Fialkov A., de Lera Acedo E., Javid K., 2021, Monthly Notices of the Royal Astronomical Society, 508, 2923–2936
Bowman J. D., Rogers A. E. E., Monsalve R. A., Mozdzen T. J., Mahesh N., 2018, Nature, 555, 67–70
Bradbury J., et al., 2018, JAX: composable transformations of Python+NumPy programs, http://github.com/jax-ml/jax
Brandenberger R., Cyr B., Shi R., 2019, Journal of Cosmology and Astroparticle Physics, 2019, 009
Cabezas A., Corenflos A., Lao J., Louf R., 2024, BlackJAX: Composable Bayesian inference in JAX (arXiv:2402.10797)
Carter G., Handley W., Ashdown M., Razavi-Ghods N., 2025, Monthly Notices of the Royal Astronomical Society, 544, 1463
Cohen A., Fialkov A., Barkana R., Lotem M., 2017, Monthly Notices of the Royal Astronomical Society, 472, 1915
Cumner J., et al., 2022, Journal of Astronomical Instrumentation, 11, 2250001
Datta A., Bradley R., Burns J. O., Harker G., Komjathy A., Lazio T. J. W., 2016, The Astrophysical Journal, 831, 6
Dawson K. S., et al., 2013, AJ, 145, 10
DeBoer D. R., et al., 2017, PASP, 129, 045001
Dowell J., Taylor G. B., Schinzel F. K., Kassim N. E., Stovall K., 2017, Monthly Notices of the Royal Astronomical Society, 469, 4537
Duane S., Kennedy A. D., Pendleton B. J., Roweth D., 1987, Physics Letters B, 195, 216
Fialkov A., Barkana R., Visbal E., Tseliakhovich D., Hirata C. M., 2013, MNRAS, 432, 2909
Fialkov A., Barkana R., Visbal E., 2014, Nature, 506, 197
Fowler J. W., et al., 2010, The Astrophysical Journal, 722, 1148
Fridman P. A., Baan W. A., 2001, A&A, 378, 327
Furlanetto S. R., Peng Oh S., Briggs F. H., 2006, Physics Reports, 433, 181–301
Gardner J. P., et al., 2006, Space Sci. Rev., 123, 485
Gessey-Jones T., Handley W. J., 2024, Phys. Rev. D, 109, 123541
Gessey-Jones T., et al., 2022, Monthly Notices of the Royal Astronomical Society, 516, 841
Gessey-Jones T., Pochinda S., Bevins H. T. J., Fialkov A., Handley W. J., de Lera Acedo E., Singh S., Barkana R., 2024, Monthly Notices of the Royal Astronomical Society, 529, 519
Handley W. J., Hobson M. P., Lasenby A. N., 2015, Monthly Notices of the Royal Astronomical Society, 453, 4384
Harker G. J. A., Pritchard J. R., Burns J. O., Bowman J. D., 2012, MNRAS, 419, 1070
Hills R., Kulkarni G., Meerburg P. D., Puchwein E., 2018, Nature, 564, E32–E34
Hoffman M. D., Gelman A., 2014, Journal of Machine Learning Research, 15, 1593
Hutter, Anne Cueto, Elie R. Dayal, Pratika Gottlöber, Stefan Trebitsch, Maxime Yepes, Gustavo 2025, A&A, 694, A254
Jeffrey N., Wandelt B. D., 2024, Machine Learning: Science and Technology, 5, 015008
Kass R. E., Raftery A. E., 1995, J. Am. Statist. Assoc., 90, 773

Kravtsov A., Belokurov V., 2024, Stochastic star formation and the abundance of $z > 10$ UV-bright galaxies (arXiv:2405.04578), https://arxiv.org/abs/2405.04578
Leeney S. A. K., 2025, JAX-bandflux: differentiable supernovae SALT modelling for cosmological analysis on GPUs (arXiv:2504.08081), https://arxiv.org/abs/2504.08081
Leeney S. A. K., Handley W. J., Acedo E. d. L., 2023, Phys. Rev. D, 108, 062006
Leeney S. A. K., Handley W. J., Bevins H. T. J., de Lera Acedo E., 2025, Bayesian Anomaly Detection for Ia Cosmology: Automating SALT3 Data Curation (arXiv:2509.13394), https://arxiv.org/abs/2509.13394
Leeney S., Meerburg P. D., Weniger C., Acedo E. d. L., Handley W., 2026, RAS Techniques and Instruments, 3, 724–736
Lemos P., Malkin N., Handley W., Bengio Y., Hezaveh Y., Perreault-Levasseur L., 2024, in Salakhutdinov R., Kolter Z., Heller K., Weller A., Oliver N., Scarlett J., Berkenkamp F., eds, Vol. 235, Proceedings of the 41st International Conference on Machine Learning. PMLR, pp 27230–27253, https://icml.cc/
Li, Zhaozhou Dekel, Avishai Sarkar, Kartick C. Aung, Han Giavalisco, Mauro Mandelker, Nir Tacchella, Sandro 2024, A&A, 690, A108
Liu H., Outmezguine N. J., Redigolo D., Volansky T., 2019, Phys. Rev. D, 100, 123011
Lovick T., Yallup D., Piras D., Mancini A. S., Handley W., 2025, High-Dimensional Bayesian Model Comparison in Cosmology with GPU-accelerated Nested Sampling and Neural Emulators (arXiv:2509.13307), https://arxiv.org/abs/2509.13307
MacKay D. J. C., 1992, Neural Computation, 4, 415
Mellema G., Koopmans L., Shukla H., Datta K. K., Mesinger A., Majumdar S., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14). p. 10 (arXiv:1501.04203), doi:10.22323/1.215.0010
Mesinger A., 2019, The Cosmic 21-cm Revolution; Charting the first billion years of our universe, doi:10.1088/2514-3433/ab4a73.
Mittal S., Ray A., Kulkarni G., Dasgupta B., 2022, Journal of Cosmology and Astroparticle Physics, 2022, 030
Mittal S., Kulkarni G., Anstey D., de Lera Acedo E., 2024, Monthly Notices of the Royal Astronomical Society, 534, 1317
Monsalve R. A., et al., 2024, The Astrophysical Journal, 961, 56
Morales M. F., Bowman J. D., Hewitt J. N., 2006, The Astrophysical Journal, 648, 767
NVIDIA Corporation 2020, Whitepaper V1.0, NVIDIA A100 Tensor Core GPU Architecture, https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf. NVIDIA, https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf
Neal R. M., 1996, Bayesian Learning for Neural Networks. Lecture Notes in Statistics Vol. 118, Springer, New York, NY, doi:10.1007/978-1-4612-0745-0
Owens J. D., Houston M., Luebke D., Green S., Stone J. E., Phillips J. C., 2008, Proceedings of the IEEE, 96, 879
Pagano M., Sims P., Liu A., Anstey D., Handley W., de Lera Acedo E., 2023, Monthly Notices of the Royal Astronomical Society, 527, 5649–5667
Pattison J. H. N., Anstey D. J., de Lera Acedo E., 2023, Monthly Notices of the Royal Astronomical Society, 527, 2413
Pattison J. H. N., Cavillot J., Bevins H. T. J., Anstey D. J., Cumner J. M., de Lera Acedo E., 2025, Monthly Notices of the Royal Astronomical Society, 538, 1301
Philip L., et al., 2018, Journal of Astronomical Instrumentation, 08
Planck Collaboration et al., 2014, A&A, 571, A16
Planck Collaboration et al., 2016, A&A, 594, A13
Planck Collaboration et al., 2020, A&A, 641, A6
Prathaban M., Yallup D., Alvey J., Yang M., Templeton W., Handley W., 2025, Gravitational-wave inference at GPU speed: A bilby-like nested sampling kernel within blackjax-ns (arXiv:2509.04336), https://arxiv.org/abs/2509.04336
Pritchard J. R., Loeb A., 2012, Reports on Progress in Physics, 75, 086901
Reis I., Fialkov A., Barkana R., 2021, Monthly Notices of the Royal Astronomical Society, 506, 5479
Sabne A., 2020, XLA : Compiling Machine Learning for Peak Performance
Sartorio N. S., et al., 2023, Monthly Notices of the Royal Astronomical Society, 521, 4039
Saxena A., Meerburg P. D., Weniger C., Acedo E. d. L., Handley W., 2024, RAS Techniques and Instruments, 3, 724–736
Schauer A. T. P., Liu B., Bromm V., 2019, ApJ, 877, L5
Scheutwinkel K. H., de Lera Acedo E., Handley W., 2022a, Publications of the Astronomical Society of Australia, 39
Scheutwinkel K. H., de Lera Acedo E., Handley W., 2022b, Publications of the Astronomical Society of Australia, 39, e052
Shaver P. A., Windhorst R. A., Madau P., de Bruyn A. G., 1999, A&A, 345, 380
Shen E., Anstey D., de Lera Acedo E., Fialkov A., Handley W., 2021, Monthly Notices of the Royal Astronomical Society, 503, 344
Shen E., Anstey D., de Lera Acedo E., Fialkov A., 2022, Monthly Notices of the Royal Astronomical Society, 515, 4565
Sims P. H., Pober J. C., 2019, Monthly Notices of the Royal Astronomical Society, 492, 22
Sims P. H., et al., 2025a, Monthly Notices of the Royal Astronomical Society, 541, 2262–2281
Sims P. H., et al., 2025b, MNRAS, 544, 2340
Singh S., Subrahmanyan R., 2019, The Astrophysical Journal, 880, 26
Singh S., et al., 2018, ApJ, 858, 54
Singh S., et al., 2022, Nature Astronomy, 6, 607
Skilling J., 2006, Bayesian Analysis, 1, 833
Smoot G. F., et al., 1992, ApJ, 396, L1
Tingay S. J., et al., 2013, Publ. Astron. Soc. Australia, 30, e007
Whitler L., et al., 2025, The $z \gtrsim 9$ galaxy UV luminosity function from the JWST Advanced Deep Extragalactic Survey: insights into early galaxy evolution and reionization (arXiv:2501.00984), https://arxiv.org/abs/2501.00984
Yallup D., Kroupa N., Handley W., 2025a, in Frontiers in Probabilistic Inference: Learning meets Sampling. https://openreview.net/forum?id=ekbkMSuPo4
Yallup D., Prathaban M., Alvey J., Handley W., 2025b, Parallel Nested Slice Sampling for Gravitational Wave Parameter Estimation (arXiv:2509.24949), https://arxiv.org/abs/2509.24949
Zheng H., et al., 2016, Monthly Notices of the Royal Astronomical Society, 464, 3486–3497
de Lera Acedo E., et al., 2022, Nature Astronomy, 6, 984–998
de Oliveira-Costa A., Tegmark M., Gaensler B. M., Jonas J., Landecker T. L., Reich P., 2008, MNRAS, 388, 247
van Haarlem, M. P. et al., 2013, A&A, 556, A2

APPENDIX A: ACCELERATION BENCHMARKING

In subsection 2.4, we described how nested sampling can be accelerated using modern GPU hardware through parallelisation at two distinct levels: within the likelihood evaluation and through algorithmic reformulation. This section benchmarks the performance gains associated with each stage by comparing the GPU-accelerated implementations with traditional CPU-based approaches, focusing on reductions in wall-clock time and overall financial cost.

A1 Likelihood Acceleration

We first evaluate the execution time of the likelihood function on both CPU and GPU architectures, examining how performance scales with the dimensionality of the parameter space (controlled by the number of regions) and with data volume. When investigating the former, to isolate the independent contributions of speed-up from compiler optimisation and hardware parallelism, we initially benchmark the effect of just-in-time (JIT) compilation on CPU execution, before

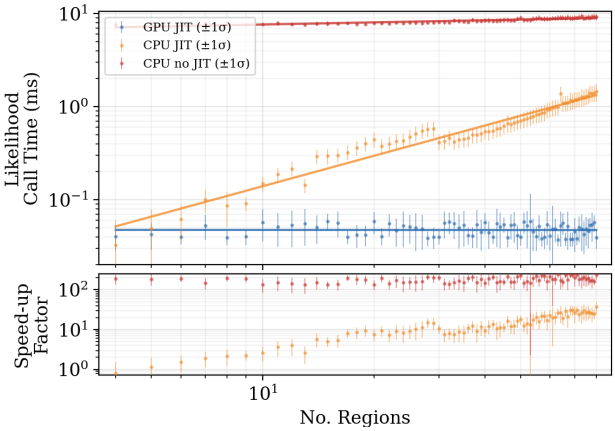

Figure A1. Performance benchmarking of the likelihood evaluation across varying model complexities. Top: Comparison of mean execution time (ms) for 1000 likelihood calls on an Intel Cascade Lake CPU (with and without JIT compilation) versus an NVIDIA A100 GPU. Bottom: The resulting speed-up factor of the A100 implementation relative to both JIT and non-JIT CPU baselines.

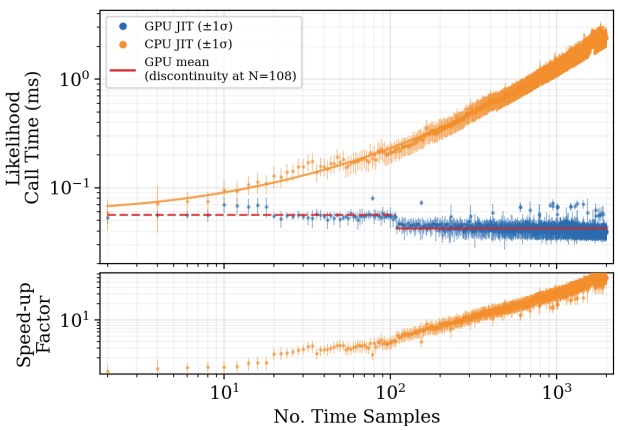

Figure A2. Performance benchmarking of the likelihood evaluation across varying data volumes. Top: Comparison of mean execution time (ms) for 1000 likelihood calls on an Intel Cascade Lake CPU (with JIT compilation) versus an NVIDIA A100 GPU, with discontinuity highlighted in by red line (dashed/solid). Bottom: The resulting speed-up factor of the A100 implementation relative to the JIT CPU baselines.

showing the maximal effect on an NVIDIA A100 GPU. The two distinct stages of computational efficiency are evident from the resulting performance trends shown in Figure A1.

JIT compilation alone yields an improvement of approximately two orders of magnitude in the constant computational overhead of the likelihood evaluation. Furthermore, while both the compiled and uncompiled CPU implementations scale approximately linearly with model dimensionality, $O(N)$, JIT compilation significantly reduces the magnitude of this scaling, from an increase of 0.023 ms per additional region in the uncompiled case to 0.017 ms per region for the compiled. This demonstrates that compiler optimisation substantially accelerates the sequential CPU execution, both by reducing fixed overheads and by improving scaling behaviour.

Moving beyond CPU execution, the GPU implementation exhibits effectively constant run-time with model dimensionality, $O(1)$, indicating near-perfect parallelisation of the likelihood evaluation with increasing parameters. Although this behaviour will ultimately be limited by available device memory, these results demonstrate that even for models with substantially inflated region counts, well beyond those required for realistic analyses, the memory capacity of a modern accelerator such as the NVIDIA A100 is sufficient.

For scaling with data volume, the effect of JIT execution is not shown explicitly at the data scales investigated (up to 2000 spectra, representative of a typical six-month REACH telescope observing window) as the computational overhead of non-compiled CPU execution renders such benchmarking infeasible. It is important to note, however, that this represents how joint, time-resolved fits at these scales was entirely impractical within the traditional CPU-based pipeline and therefore the current framework is essential for processing the full volume of observational data. We therefore focus on the comparative scaling behavior of the JIT-compiled CPU and GPU implementations using a constant 10-region model, as shown in Figure A2. Once again we show that the sequential processing nature of CPUs leads to increased runtime with data volume in comparison to $O(1)$ scaling on a GPU.

One noticeable trend in the GPU runtime is a discontinuity at 108 time samples, resulting in a $\approx 20\,\%$ reduction in runtime. We attribute this to the hardware specifications of the NVIDIA A100 (80GB), which features 108 Streaming Multiprocessors (SMs). Therefore as the operations achieve sufficient occupancy to allow for Tensor Core acceleration, a transition in the underlying execution strategy occurs, moving the workload from standard Single Instruction, Multiple Thread (SIMT) execution on the FP64 CUDA cores (9.7 TFLOPS) to the more efficient Single Instruction, Multiple Data (SIMD) style processing of the FP64 Tensor Cores (19.5 TFLOPS) (NVIDIA Corporation 2020).

A2 Nested Sampling Acceleration

Finally, we benchmark the total execution time of the `BlackJAX` algorithm employed in this work against a traditional CPU-based Nested Sampler, taken in this work to be `PolyChord` (Handley et al. 2015). This comparison evaluates the end-to-end performance of the inference pipeline, encompassing both the sampler's algorithmic efficiency and the accelerated likelihood evaluations. For fair comparison, we perform these runs with consistent algorithmic configurations to match the defaults of `PolyChord`: `n_live` $= 25 \times$ `nDim` and `num_inner_steps` $= 5 \times$ `nDim` (see section B for more details).

APPENDIX B: ALGORITHMIC CONVERGENCE

The hyperparameters for the GPU-accelerated Nested Sampling runs presented in subsubsection 2.4.3 were selected to balance computational efficiency with algorithmic precision. While the high computational cost of traditional CPU-based Nested Sampling often necessitates approximate hyperparameter choices, the new accelerated framework allows for a systematic exploration of convergence.

To establish a robust parameter configuration, we performed a convergence study on a 23-parameter model (19 linearly split regions) using the Galaxy Up observational data. This configuration was chosen as an extreme test case, as its dimensionality exceeds that of the primary analysis. As shown in Figure B1, while the consistency of the Bayesian evidence ($\ln \mathcal{Z}$) is influenced by both the density of live points (`n_live`) and the number of inner slice-sampling steps (`num_inner_steps`), the latter exhibits a significantly greater effect. This indicates that the primary driver of stability for this case

Configuration	No. Regions	GPU Runtime (s)	CPU Runtime (s)	Speed-up Factor	Price Factor
Integrated	10	25.48	—	—	—
Integrated	20	39.68	—	—	—
Integrated	30	72.71	—	—	—
Resolved	10	39.04	—	—	—
Resolved	20	123.19	—	—	—
Resolved	30	315.13	—	—	—

Table A1. Comparison of runtime and cost efficiency between `BlackJAX` executed on an NVIDIA A100 GPU and `PolyChord` executed on a 40-core Intel Ice Lake CPU. Runtimes are evaluated for a fixed six-hour observation window corresponding to 72 observations. The speed-up factor is defined as $t_{\mathrm{CPU}}/t_{\mathrm{GPU}}$, while the price factor denotes the relative financial cost per run based on hardware pricing on the Cambridge Service for Data Driven Discovery (CSD3) HPC system, assuming costs of 0.01 p per CPU core-hour and 100 p per A100 GPU-hour.

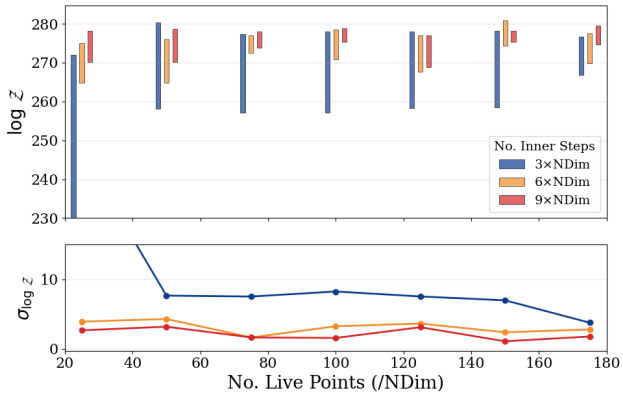

Figure B1. Consistency of Bayesian evidence estimates from BlackJAX nested sampling for the Galaxy-Up 1-hour integration with 19 linearly split regions. The top panel shows the spread of recovered $\ln \mathcal{Z}$ across repeats, and the bottom panel shows $\sigma_{\log \mathcal{Z}}$, both as functions of the number of live points (per N_{Dim}). Colours indicate the number of inner slice-sampling steps (3, 6, or $9\times N_{\mathrm{Dim}}$). All runs use a `num_delete` of $0.2 \times N_{\mathrm{Dim}}$.

is the effective decorrelation of samples and thorough exploration of the prior volume. Consequently, all results in this work use an even more conservative value of `num_inner_steps` = $12 \times$ `nDim` to ensure reliable and repeatable evidence estimation across all observing windows.

Furthermore, the efficiency of this framework enables all future analyses performed on true observational data to apply the same bootstrapping procedures to verify that convergence is universal across the entire dataset and all model configurations.

This paper has been typeset from a TEX/LATEX file prepared by the author.