

Constraining reionization morphology and source properties with 21cm galaxy cross-correlation surveys

Yannic Pietschke^{1,2,*}, Anne Hutter², and Caroline Heneka¹

¹ Institut für Theoretische Physik, Universität Heidelberg, Philosophenweg 16, 69120 Heidelberg, Germany

² Institute for Astronomy, University of Vienna, Türkenschanzstrasse 17, A-1180 Vienna, Austria

Received - -, -; accepted - -, -

ABSTRACT

Cross-correlations between 21cm observations and galaxy surveys provide a powerful probe of reionization by providing robustness against foreground contamination while linking ionization morphology to galaxies. We quantified the constraining power of 21cm galaxy cross-power spectra for inferring the neutral hydrogen fraction, $x_{\text{HI}}(z)$, and mean overdensity, $\langle 1 + \delta_{\text{HI}} \rangle(z)$, exploring dependence on the field of view; redshift precision, σ_z ; and minimum halo mass, $M_{\text{h,min}}$. We employed our simulation-based inference framework EoRFlow for likelihood-free parameter estimation. Mock observations include thermal noise for 100h of SKA-Low with foreground avoidance and realistic galaxy-survey effects. For a fiducial survey (FOV = 100 deg², $\sigma_z = 0.001$, $M_{\text{h,min}} = 10^{11} M_{\odot}$), cross-power spectra yield unbiased constraints with posterior volumes (PVs) of $\sim 10\%$ relative to priors. Cross-power measurements reduce the PV by 20–30% versus 21cm auto-power alone. With foreground avoidance, spectroscopic redshift precision is essential; photometric redshifts render cross-correlations uninformative. Notably, cross-power spectra constrain ionizing source properties, the escape fraction f_{esc} , and the star formation efficiency f_* , which remain degenerate in auto-power (PV >60%). Tight constraints require either deep surveys detecting faint galaxies ($M_{\text{h,min}} \sim 10^{10} M_{\odot}$) with moderate foregrounds (PV $\sim 11\%$) or conservative mass limits with optimistic foreground removal (PV $\sim 19\%$). 21cm galaxy cross-correlations enhance morphology constraints beyond auto-power while enabling previously inaccessible source property constraints. Realizing full potential requires precise redshifts and either faint galaxy detection limits or improved 21cm foreground cleaning.

Key words. Galaxies: high-redshift – intergalactic medium – dark ages, reionization, first stars – large-scale structure of Universe – Methods: numerical, statistical

1. Introduction

The epoch of reionization (EoR) marks the transformational period in cosmic history when the first generations of galaxies and quasars emitted radiation that ionized the intergalactic medium (IGM). Observations of the redshifted 21cm hyperfine transition of neutral hydrogen offer a novel 3D view of this era, encoding the distribution of ionized and neutral regions across cosmic time. Low-frequency radio interferometers such as the Hydrogen Epoch of Reionization Array (HERA), the Murchison Wide-field Array (MWA), the LOw-Frequency ARray (LOFAR), and the Square Kilometre Array (SKA)¹ are actively pursuing this signal through increasingly sensitive observations (Parsons et al. 2010; Tingay et al. 2013; DeBoer et al. 2017; van Haarlem, M. P. et al. 2013; Mertens et al. 2020; Trott et al. 2020; Yoshiura et al. 2021; Abdurashidova et al. 2022, 2023; Ceccotti et al. 2025). However, extracting robust astrophysical constraints from 21cm measurements remains extremely challenging due to the high dimensionality of the signal, non-Gaussianity, and the significant contamination from astrophysical foregrounds several orders of magnitude brighter than the cosmological signal.

Cross-correlations between 21cm observations and high-redshift galaxy surveys provide a promising avenue to both confirm the cosmological origin of the 21cm signal and enhance constraints on reionization models (Furlanetto & Lidz 2007; Park et al. 2014; Sobacchi et al. 2016; Vrbanc et al. 2016;

Heneka et al. 2017; Hutter et al. 2017; Kubota et al. 2018; Yoshiura et al. 2018; Beane et al. 2019; Heneka & Mesinger 2020; Vrbanc et al. 2020; Weinberger et al. 2020; Heneka & Cooray 2021; La Plante et al. 2023; Moriwaki et al. 2024; Gagnon-Hartman et al. 2025; Hutter & Heneka 2026). Because foregrounds in radio interferometry are largely uncorrelated with galaxy positions, the 21cm galaxy cross-power spectrum is unbiased by foreground contamination (Furlanetto & Lidz 2007; Lidz et al. 2009; Yoshiura et al. 2018) and links the evolving ionization morphology to observable galaxy populations. Recent work has developed realistic forecasts for detecting the 21cm galaxy cross-power spectrum across instrument configurations and survey parameters, identifying regimes where signal-to-noise ratios (S/N) are maximized and where differing ionizing source scenarios can be distinguished observationally (Gagnon-Hartman et al. 2025; Hutter & Heneka 2026). These studies highlight how cross-power measurements can trace the evolving morphology of ionized regions and advance individual 21cm or galaxy-only probes of reionization. Upcoming facilities such as the Prime Focus Spectrograph (PFS) on Subaru (Greene et al. 2022), the Nancy Grace Roman Space Telescope (Wang et al. 2022), the MOONRISE program (MOONS Extragalactic GTO; Maiolino et al. 2020), and the proposed Widefield Spectroscopic Telescope (WST; Mainieri et al. 2024) will provide high-redshift galaxy samples spanning a range of survey areas, depths, and redshift precision, making practical survey optimization a pressing concern for maximizing scientific return from cross-correlation measurements.

* e-mail: pietschke@thphys.uni-heidelberg.de

¹ <https://www.skao.int/en>

With these diverse survey capabilities on the horizon, key questions remain regarding what astrophysical information cross-correlations can robustly extract and how survey design impacts scientific return. The neutral hydrogen fraction, $x_{\text{HI}}(z)$, and the mean overdensity in neutral regions, $\langle 1 + \delta_{\text{HI}} \rangle(z)$, together encode both the EoR timeline and the ionization morphology (e.g., the distinction between reionization scenarios driven either by low-mass or massive galaxies) (Hutter et al. 2023a). However, whether cross-power measurements can constrain these quantities under realistic noise and survey limitations remains uncertain. Furthermore, source properties such as the escape fraction of hydrogen-ionizing photons from galaxies and star formation efficiency—which fundamentally drive reionization—remain poorly constrained observationally, raising the question of whether cross-correlations offer a pathway to access these otherwise inaccessible parameters.

Addressing these questions requires robust statistical frameworks that can handle the intrinsic complexity of the joint data and the intractability of traditional likelihoods. The high-dimensional, non-Gaussian nature of the 21cm signal, combined with the absence of analytically accurate forward models relating astrophysical parameters to observables, renders traditional likelihood-based inference computationally prohibitive or impossible. Simulation-based inference (SBI), also known as likelihood-free inference, has recently emerged as a powerful tool in cosmological analyses (Alsing et al. 2019; Cole et al. 2022; Villaescusa-Navarro et al. 2022; Saxena et al. 2024; Schosser et al. 2025; Ore et al. 2025), including the joint analysis of multiple probes (Schosser et al. 2026), circumventing explicit likelihood modeling by learning the mapping from observables to parameters directly from forward simulations. Our framework *EoRFlow*² (Pietschke et al. 2025), originally trained on 21cm power spectra from diverse reionization models, exemplifies this approach. *EoRFlow* uses neural density estimators to efficiently and unbiasedly reconstruct the evolution of the global neutral hydrogen fraction $x_{\text{HI}}(z)$ without relying on approximate likelihoods. This method enables scalable posterior estimation across narrow redshift slices and has been validated on realistic mock datasets incorporating instrument noise, demonstrating its potential for reconstructing reionization histories from forthcoming SKA-Low observations.

In this work, we extended the SBI paradigm to the 21cm galaxy cross-power spectrum, combining the realistic cross-power modeling developed in Hutter & Heneka (2026) with the flexible inference power of *EoRFlow*. By integrating simulated cross-power spectra with comprehensive noise and survey models into an SBI pipeline, we inferred the aforementioned astrophysical parameters governing reionization, directly linking the statistical imprint of high-redshift galaxies and neutral hydrogen fluctuations. This approach is essential because the absence of tractable analytic models connecting astrophysical parameters to cross-power observables makes SBI the natural framework for extracting physical information from these complex, high-dimensional measurements. Tackling cross-correlations with SBI not only enhances the robustness of parameter constraints by exploiting complementary information from the cross-power, it also mitigates the need for analytically tractable likelihoods in high-dimensional regimes where traditional methods struggle.

The structure of this paper is as follows. We begin with a description of our simulation framework and mock data generation in Section 2. Section 3 then details the neural-network architecture and training procedures. Subsequently, Section 4 presents

our main results. We start with the inference of global reionization properties (Section 4.1). We then systematically investigate the impact of galaxy survey parameters (Section 4.2) and the information gained from cross-correlation measurements (Section 4.3). Moreover, we demonstrate the capabilities of the cross-power spectrum to infer reionization source properties (Section 4.4). Finally, we conclude in Section 5.

2. The 21cm galaxy mock data

We generated mock realizations of the 21cm signal and the underlying galaxy population using 21cmFASTv4 (Mesinger et al. 2011; Murray et al. 2020; Park et al. 2019; Davies et al. 2025). The simulations produced coeval 3D cubes of halo properties and the corresponding differential 21cm brightness temperature, δT_{b} , during the EoR.

We simulated coeval volumes at the redshifts $z \in [6.0, 6.3, 6.6, 7.0, 7.3, 7.6, 8.0]$, which span the late stages of reionization where current and forthcoming galaxy surveys, including PFS, Roman, MOONRISE, and WST, will enable the detection of 21cm galaxy cross-correlations through high-redshift galaxy samples. At these redshifts ($z \leq 8$), we assumed a saturated spin temperature, as expected in standard EoR scenarios, corresponding to the post-heating regime in which $T_{\text{S}} \gg T_{\text{CMB}}$, and consistent with current upper limits from HERA (Abdurashidova et al. 2022; HERA Collaboration et al. 2023; Abdurashidova et al. 2026). Under this assumption, the 21cm brightness temperature depends only on the neutral hydrogen fraction and the density field via (Furlanetto et al. 2006)

$$\delta T_{\text{b}}(\mathbf{x}) \simeq T_0 x_{\text{HI}}(\mathbf{x})(1 + \delta(\mathbf{x})). \quad (1)$$

All simulations were performed in a cubic volume of side length 300 cMpc with a cell size of 2 cMpc, providing sufficient resolution to capture the large-scale ionization morphology relevant for cross-correlation studies while maintaining computational efficiency.

To capture a broad range of physically plausible reionization scenarios, we varied the ionizing photon escape fraction, f_{esc} , and the star formation efficiency, f_* , according to the power-law parameterization implemented in 21cmFASTv4 (Park et al. 2019):

$$f_*(M_{\text{h}}) = f_{*,10} \left(\frac{M_{\text{h}}}{10^{10} M_{\odot}} \right)^{\alpha_*}, \quad (2)$$

$$f_{\text{esc}}(M_{\text{h}}) = f_{\text{esc},10} \left(\frac{M_{\text{h}}}{10^{10} M_{\odot}} \right)^{\alpha_{\text{esc}}}, \quad (3)$$

where $f_{*,10}$ and $f_{\text{esc},10}$, as well as α_* and α_{esc} , are the normalization and power-law exponents of the star formation efficiency and escape fraction, respectively. We fixed other astrophysical and cosmological parameters to fiducial values consistent with Planck 2018 results ($\Omega_{\text{b}} = 0.049$, $\Omega_{\text{m}} = 0.31$, $h = 0.68$; Aghanim et al. 2020) and adopted the standard 21cmFAST prescriptions for stellar spectra and X-ray heating. To exclude scenarios of extremely delayed or early reionization, we filtered out models where the neutral fraction remains either $x_{\text{HI}} > 0.7$ or $x_{\text{HI}} < 0.2$ across all redshifts in our observational window ($z = 6.0$ - 8.0). The simulation parameters and their prior ranges are summarized in Table 1. Uniform priors are assumed over the specified intervals. These parameters jointly control the abundance and clustering of ionizing sources and therefore strongly influence both the morphology of ionized regions and the amplitude and scale dependence of the 21cm galaxy cross-power

² <https://github.com/astro-ML/EoRFlow>

Table 1: Summary of simulation parameters and prior ranges.

Parameter	Prior Range
$\log_{10} f_{\text{esc},10}$	$\mathcal{U}[0.005, 0.5]$
α_{esc}	$\mathcal{U}[-0.8, 0.5]$
$\log_{10} f_{*,10}$	$\mathcal{U}[0.005, 0.5]$
α_*	$\mathcal{U}[-0.3, 0.9]$

spectrum. In addition to these astrophysical parameters, we also varied the random seed for each simulation to account for cosmic variance.

To create realistic mock observations of the 21cm auto-power spectrum and the 21cm galaxy cross-power spectrum, we employed the forward-modeling framework developed in Hutter & Heneka (2026). This method incorporates instrumental noise, survey geometry, and galaxy-selection effects in a self-consistent manner, enabling direct comparison with realistic observational scenarios. From each simulated coeval cube, we computed the spherically averaged 1D power spectra $P_{21}(k)$ and $P_{21,g}(k)$. For the galaxy field, halos were selected according to survey-specific detection thresholds, and their spatial distribution was used to construct the galaxy overdensity field. The cross-power spectrum is then computed by correlating the galaxy overdensity with the 21cm brightness temperature fluctuations. Here, we used 15 linearly spaced bins with $k \in [0.08, 2.40]$. Thermal noise contributions to the 21cm measurements and shot noise in the galaxy field were added following the prescriptions described in Hutter & Heneka (2026). The 21cm noise calculations are based on the public code 21cmSense (Poher et al. 2013, 2014). We assumed 100h of SKA AA* observations as well as the moderate foreground model in which the foreground wedge extends 0.1Mpc^{-1} beyond the horizon limit. For the fiducial galaxy survey configuration used in Section 4.1, we adopted a field of view (FOV) of 100deg^2 , spectroscopic redshift precision $\sigma_z = 0.001$, and a minimum halo mass of $M_{\text{h,min}} = 10^{11} M_{\odot}$, representing an optimistic but realistic large-area spectroscopic survey, as proposed for the WST. In Section 4.2 we systematically explore the impact of varying these parameters. For the largest FOV considered here (100deg^2), the signal power is computed from the simulation box (side length 300cMpc), while the uncertainties are scaled to the larger FOV by accounting for the increased number of k -modes.

Note that in this work we treated halos and their attributed masses as proxies for galaxies. For cross-correlation studies, emission-line galaxy surveys—such as Lyman- α and [OIII] emitters—are ideal due to their precise redshift measurements. To the first order, the luminosities of these emission lines scale with the star formation rate, which increases with halo mass, justifying our use of halo mass as a proxy for galaxy detectability. Accordingly, halo masses can be translated to their corresponding line luminosities through empirical models. For example, halo masses of $M_{\text{h}} = 10^{10}$, $10^{10.5}$, and $10^{11} M_{\odot}$ correspond to Lyman- α luminosities of $L_{\text{Ly}\alpha} \approx 10^{41}$, $10^{41.8}$, and $10^{42.3} \text{erg s}^{-1}$, respectively, at $z \sim 7$ (see Fig. 5 in Hutter et al. 2023b).

To generate mock observations of the power spectra that reflect the combined impact of astrophysical modeling and observational limitations, we sampled from a normal distribution of $\mathcal{N}(P_{21,g}(k), \sigma_{21,g}(k))$, where $P_{21,g}(k)$ is the true, physical cross-power spectrum and $\sigma_{21,g}(k)$ represents the observational uncertainties determined by the FOV; redshift precision, σ_z ; and minimum halo mass, $M_{\text{h,min}}$; as well as thermal noise from SKA-Low AA* (100h) and foreground avoidance. An example cross-power spectrum is shown in Fig. 1. In total, we produced 10790 sam-

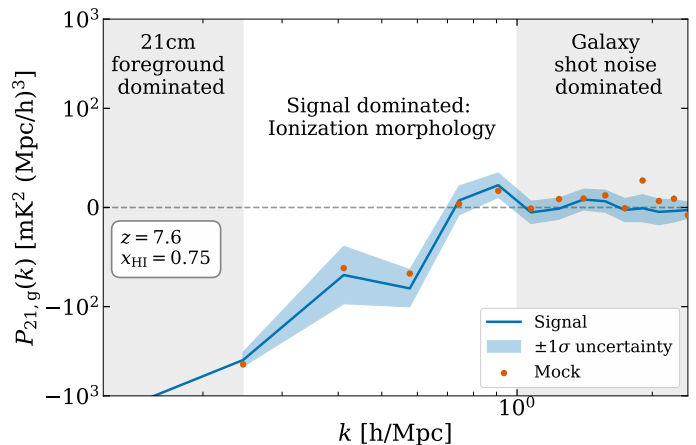


Fig. 1: The 21cm galaxy cross-power spectrum at $z = 7.6$ ($x_{\text{HI}} = 0.75$) for the fiducial survey configuration explored in the inference analysis in Section 4.1 (FOV = 100deg^2 , $\sigma_z = 0.001$, $M_{\text{h,min}} = 10^{11} M_{\odot}$). The blue line shows the physical signal, the shaded blue region indicates the 1σ uncertainty, and orange points represent the mock observation including instrumental noise. Three distinct regimes characterize the cross-power spectrum. The first region corresponds to the 21cm foreground wedge where large-scale modes are contaminated by bright foreground emission, resulting in poor S/N. We omitted the signal here based on our foreground-avoidance strategy. The second region is signal dominated and provides access to the ionization morphology. It exhibits strong anticorrelation on intermediate scales ($k \sim 0.25\text{--}0.75 h\text{Mpc}^{-1}$) where galaxies predominantly occupy ionized bubbles, while the 21cm signal traces neutral hydrogen in the IGM. At $k \sim 0.75 h\text{Mpc}^{-1}$, the zero crossing corresponds to the characteristic scale of ionized regions. On scales smaller than the typical size of ionized bubbles, both the galaxy distribution and neutral hydrogen density respond to the same underlying matter fluctuations, resulting in positive correlation. The third region shows the shot-noise-dominated regime on small scales ($k > 1.0 h\text{Mpc}^{-1}$) owing to the discrete nature of galaxy datasets.

ples (after filtering) with varying astrophysical parameters and random seed.

One of our goals is to infer the full posterior probability distribution of neutral hydrogen fractions $x_{\text{HI}}(z)$ and the mean overdensity in neutral regions, $\langle 1 + \delta_{\text{HI}} \rangle(z)$, at different redshifts, z from this mock data. The latter can be obtained from Eq.(1). At a given redshift we computed the mean differential brightness temperature, $\langle \delta T_{\text{b}} \rangle$, and neutral fraction, \bar{x}_{HI} . For a standard cosmology with $\Omega_{\text{b}} = 0.049$, $\Omega_{\text{m}} = 0.31$ and $h = 0.68$, the normalization factor is given by $T_0 = 27 \sqrt{0.1(1+z)} \text{mK}$ (Mesinger et al. 2011), from which the average density can then be calculated as

$$\langle 1 + \delta_{\text{HI}} \rangle(z) = \frac{\langle \delta T_{\text{b}} \rangle(z)}{27 \sqrt{0.1(1+z)} \bar{x}_{\text{HI}}(z) \text{mK}}. \quad (4)$$

Here, $\bar{x}_{\text{HI}}(z)$ denotes the volume-averaged neutral fraction, and $\langle 1 + \delta_{\text{HI}} \rangle(z)$ is the mean density contrast averaged over neutral cells only. This separation is exact in 21cmFAST, where $x_{\text{HI}}(\mathbf{x})$ is binary for each cell. These quantities are directly computed from the 21cm simulation outputs and saved as labels for network training. This parameter combination enables constraints on both the EoR timeline and the distinction of different reionization scenarios (Hutter et al. 2023a).

3. Network architecture and training

We performed SBI using EoRFlow, a conditional normalizing flow framework originally introduced in Pietschke et al. (2025) for inference on 21cm power spectra. We also explored an inference model based on conditional flow matching (CFM; Lipman et al. 2023; Tong et al. 2024), motivated by recent successes in modeling complex distributions with more flexible continuous-time dynamics. While CFM achieves comparable posterior accuracy, posterior sampling is significantly slower (about 6x in our case) due to the need to numerically integrate an ordinary differential equation for each draw. Given the absence of clear performance gains and the substantially increased computational cost, we do not adopt CFM as our fiducial inference model.

In the present work, EoRFlow was adapted to operate on 21cm galaxy cross-power spectra, while preserving the core design philosophy of directly inferring astrophysical parameters from summary statistics derived from forward simulations, without constructing an explicit likelihood. The inference network directly took the power spectrum as input, and no additional neural summary network was employed. This choice reflects the relatively low dimensionality of the spherically averaged power spectrum and avoids introducing unnecessary architectural complexity or information bottlenecks. The network input consists of spherically averaged 1D power spectra; the 21cm auto-power, $P_{21}(k, z)$; the 21cm galaxy cross-power, $P_{21,g}(k, z)$; or both combined, evaluated in discrete redshift bins. This flexibility to seamlessly incorporate single or multiple observables is a key advantage of the SBI approach. For each redshift slice, the power spectrum is sampled on a fixed set of k-modes, forming a 1D vector that fully characterizes the observable(s) used for inference.

The inference network is implemented as a conditional normalizing flow, which models the posterior distribution $p(\theta|\mathbf{x})$, where θ denotes the astrophysical parameters of interest and \mathbf{x} the input cross-power spectrum. The overall flow structure follows Pietschke et al. (2025) and consists of a sequence of coupling transformations interleaved with permutations to ensure full mixing of parameters. However, affine coupling layers were replaced by rational quadratic spline coupling layers (Durkan et al. 2019). Spline-based coupling layers provide greater expressivity than affine transformations, allowing the flow to model highly non-Gaussian posterior distributions with fewer layers. This increased flexibility is particularly important for 21cm galaxy cross-correlation data, where the posterior structure can be influenced by complex interactions between galaxy bias, ionization morphology, and observational noise. We employed ten coupling layers with 512 nodes each and ReLU activation. All training and evaluation was performed in PyTorch (Ansel et al. 2024) with the AdamW optimizer and default decoupled weight decay for regularization (Loshchilov & Hutter 2019).

From the entire dataset of 10790 samples, we used 7832 for training, 1958 for validation, and 1000 for testing. All evaluations shown in this work were performed on the test set. For each sample the network condition consists of one power spectrum per redshift, each of which consists of 15 k-bins. For our seven redshifts this results in a total flattened array size of 105. We also appended the redshift values for temporal context, bringing the total condition dimension to 112. The power spectrum values were preprocessed by an inverse hyperbolic sine (asinh) transformation, which handles the wide dynamic range (similar to a logarithmic transformation) while accepting both positive and negative values, as present in the cross-power spectrum. Subsequently, we applied z-score normalization. Additionally,

we transformed the x_{HI} labels using a logit transformation. We trained the model with a batch size of 16 and an initial learning rate of 10^{-3} , which was halved every ten epochs without improvement in the validation loss. Early stopping was employed to avoid overfitting. Our fiducial model, trained on the cross-power spectrum for the survey configuration described in Section 2, converges after 40 epochs. When including the 21cm auto-power spectrum alongside the cross-power, training lasted for 54 epochs. On one NVIDIA A100 GPU, training took approximately 20–30 minutes, depending on survey parameters.

4. Results

In this section, we demonstrate the inference capabilities of our SBI framework by applying the trained EoRFlow to mock power-spectrum observations spanning a range of observational scenarios. Throughout this analysis, we focused on 1D power spectra. Even though the 2D cross-power spectrum should, in principle, be more sensitive to the ionization morphology and galaxy bias, we found that the higher S/N through spherical averaging surpasses this information loss. A comparison of the impact of power spectrum dimensionality can be found in Appendix A.

We organized our results into four complementary analyses. Assuming a fiducial survey configuration (see Section 2), we began by constraining the neutral hydrogen fraction, $x_{\text{HI}}(z)$, and the mean density contrast in neutral regions $\langle 1 + \delta_{\text{HI}} \rangle(z)$ across seven redshift slices (Section 4.1), comparing three training configurations: 21cm auto-power alone, cross-power alone, and both combined. For the combined model, we concatenated independently normalized power spectra into a single conditioning vector, doubling its dimensionality. While this in principle could benefit from architectural adjustments to optimally exploit the higher dimensional input, we deliberately fixed the network architecture to be identical across all three configurations to ensure fair comparison. Any performance differences thus reflect the information content of the data rather than hyperparameter tuning.

Once the baseline inference performance for our fiducial survey is established, we systematically varied galaxy survey parameters (FOV, σ_z , $M_{\text{h,min}}$) to understand which observational characteristics most strongly influence inference quality and to identify survey designs that optimize scientific return (Section 4.2). Afterwards, we employed a mutual information analysis to decompose the complementary information between auto-power and cross-power measurements on a per-bin basis, revealing which scales and redshifts benefit most from multi-tracer observations (Section 4.3). Finally, we extended our analysis beyond global quantities to infer the astrophysical properties of ionizing sources, demonstrating that cross-correlation measurements provide access to parameters that are fundamentally inaccessible to 21cm auto-power observations alone (Section 4.4).

4.1. Inferring global reionization properties

We begin by demonstrating the inference capabilities of our framework for constraining the global evolution of reionization, specifically the neutral hydrogen fraction, $x_{\text{HI}}(z)$, and the mean density contrast in neutral regions, $\langle 1 + \delta_{\text{HI}} \rangle(z)$, across seven redshift slices spanning $z = 6.0$ to 8.0 . These quantities provide complementary views of the reionization process. The neutral fraction directly quantifies the progression of reionization, while the mean density in neutral regions reflects the connection between the matter distribution and the ionization morphology, capturing how reionization proceeds preferentially in overdense regions where the first galaxies formed.

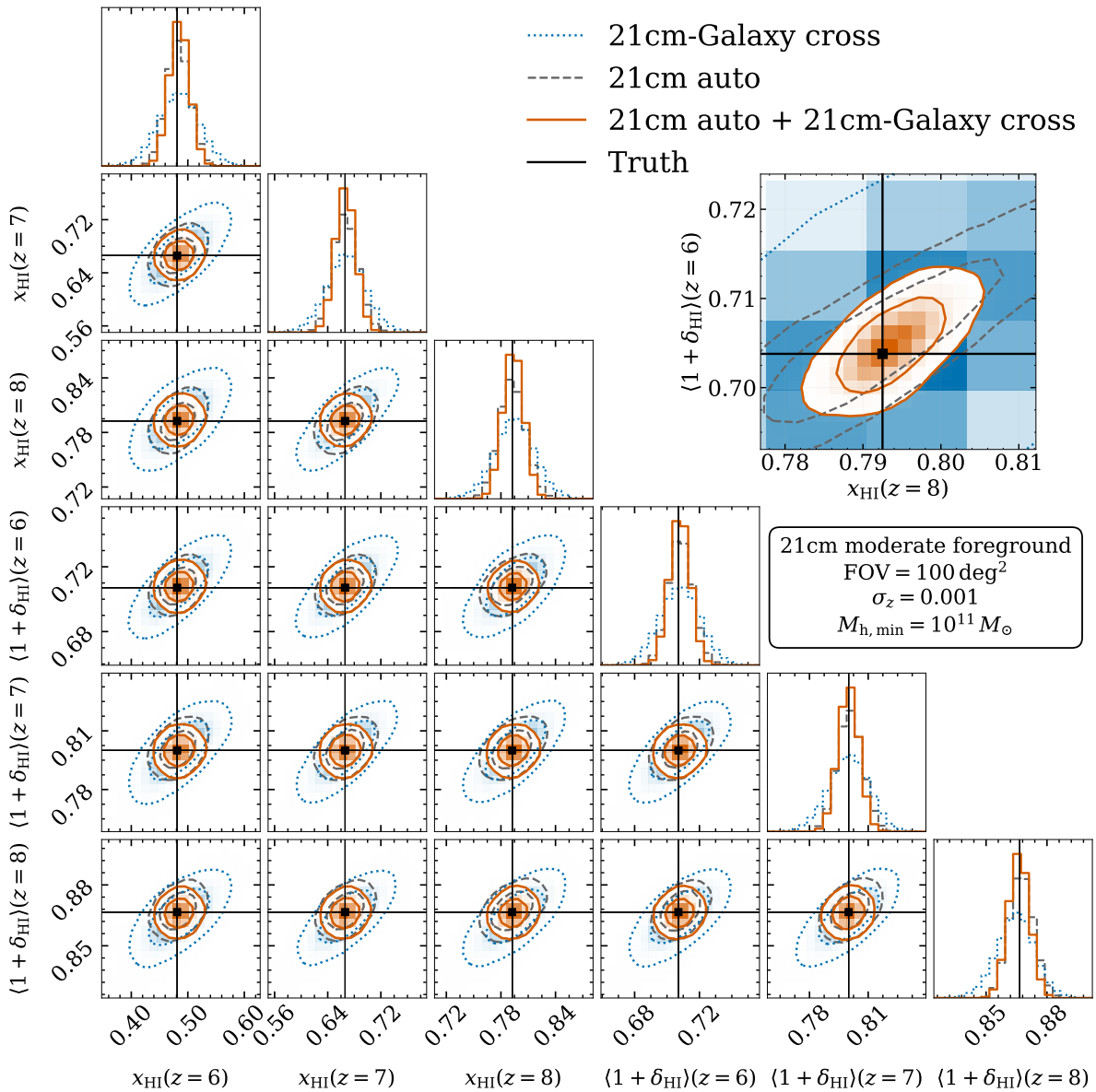


Fig. 2: Marginalized posteriors for neutral fraction, $x_{\text{HI}}(z)$, and the mean density in neutral regions, $\langle 1 + \delta_{\text{HI}} \rangle(z)$, for a selection of redshift slices and a randomly chosen set of parameters ($\log_{10} f_{*,10} = 0.04$, $\log_{10} f_{\text{esc},10} = 0.05$, $\alpha_* = 0.44$, and $\alpha_{\text{esc}} = -0.60$) assuming moderate 21cm foreground avoidance. Shadings indicate the 68% and 95% confidence intervals. Dotted blue contours show the posteriors derived from 21cm galaxy cross-power spectra, while solid orange lines represent the results when combining both cross- and auto-power spectra. The dashed gray contours denote the model trained on 21cm auto-power spectra only, and true parameter values are indicated by black dots. This realization serves as an illustrative example; posterior calibration was confirmed via coverage statistics following Pietschke et al. (2025).

For this analysis, we adopted a fiducial observational configuration that combines moderate foreground assumptions with realistic survey parameters as detailed in Section 2. This configuration represents an ambitious but achievable observational program for the coming decade, combining SKA-Low 21cm observations with wide-field spectroscopic galaxy surveys such as the WST, and it serves as our baseline for comparing different data combinations. We trained three separate conditional normalizing flow models on mock observations drawn from this fiducial survey, using power spectra from 1000 test realizations to evaluate inference performance. As described above, the three models differ only in their input data: using the 21cm auto-power spectrum alone, the cross-power spectrum alone, or both combined.

Figure 2 presents the marginalized posterior distributions for a representative test observation with randomly chosen parameters: $\log_{10} f_{*,10} = 0.04$, $\log_{10} f_{\text{esc},10} = 0.05$, $\alpha_* = 0.44$, and $\alpha_{\text{esc}} = -0.60$. The corner plot shows the inferred neutral fraction and mean density across four selected redshift slices. The 68% and 95% confidence regions reveal several key features of the inference quality. First of all, all three models recover the true parameter values (marked by black dots) within their respective uncertainty intervals for this particular realization, which was selected as a representative illustration. The calibration of our inference method was verified through coverage statistics following Pietschke et al. (2025). Moreover, the posteriors for x_{HI} are generally somewhat more tightly constrained than those for $\langle 1 + \delta_{\text{HI}} \rangle$, reflecting the fact that the neutral fraction is the pri-

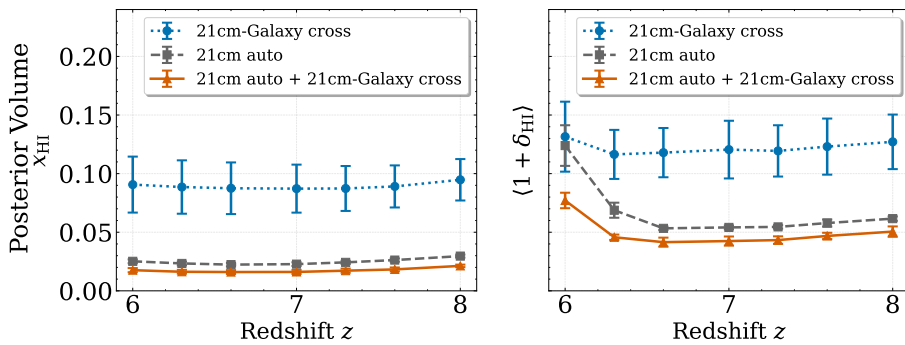


Fig. 3. Informativeness measured by the normalized posterior volume on the neutral fraction $x_{\text{HI}}(z)$ and the mean density in neutral regions $\langle 1 + \delta_{\text{HI}} \rangle(z)$ as a function of redshift assuming moderate 21cm foreground avoidance as well as our fiducial galaxy survey (100 deg², $\sigma_z = 0.001$, $M_{\text{h,min}} = 10^{11} M_{\odot}$). Points show the mean posterior volume across 6 independent network training runs and error bars indicate the standard deviation.

primary driver of power-spectrum amplitude, while density fluctuations contribute a secondary modulation. Furthermore, the relative sizes of the confidence regions immediately reveal the hierarchy of constraining power among the three data combinations, with the combined model (solid orange contours) providing the tightest constraints, followed by the 21cm auto-power model (dashed gray contours), and finally the cross-power-only model (dotted blue contours).

To quantify this hierarchy more precisely across the full test set, we computed the posterior volume (PV) for each model, which is defined as the standard deviation of the posterior samples normalized by the prior standard deviation. This metric measures what fraction of the prior uncertainty remains after observing the data, with values near unity indicating uninformative observations and values near zero indicating precise constraints. Fig. 3 shows the mean PV across six independent network training runs as a function of redshift for both x_{HI} (left panel) and $\langle 1 + \delta_{\text{HI}} \rangle$ (right panel), with error bars indicating the standard deviation. Averaging over multiple runs accounts for the epistemic uncertainty inherent to neural-network training, where different random initializations can yield slightly different posteriors; this is discussed in detail in Appendix B.

The results confirm the qualitative impression from the corner plot while revealing additional structure. For the neutral fraction, the 21cm auto-power spectrum (gray) achieves remarkably tight constraints with the PV below 0.03 across all redshifts, demonstrating the direct sensitivity of the 21cm signal to x_{HI} . The cross-power spectrum alone (blue) yields broader posteriors with a PV around 0.09, reflecting its indirect connection to the neutral fraction through the correlation between ionization bubbles and galaxies. Combining both measurements (orange) further improves constraints to the PV below 0.02, indicating that despite the strong baseline from auto-power alone, the cross-power spectrum does provide modest complementary information.

For the mean neutral hydrogen-density contrast, the pattern is similar but generally has a larger PV, particularly at lower redshifts. This redshift dependence likely reflects the decreasing size and contrast of neutral regions as reionization progresses toward completion at $z \sim 6$, making the density signal weaker and more challenging to extract. Across all redshifts, the combined model achieves a PV around 0.04, auto-power alone reaches 0.07, and cross-power alone yields 0.12. In both cases, the improvements from combining measurements are statistically significant but modest, typically yielding 20–30% reductions in posterior width compared to auto-power alone.

These findings establish several important conclusions for interpreting the remainder of our results. Firstly, 21cm auto-power spectra alone provide excellent constraints on the global reionization history, demonstrating the power of direct neutral

hydrogen observations, which confirms the findings of previous works (Pietschke et al. 2025; Cooper et al. 2026; Cerardi et al. 2025). Furthermore, cross-power measurements do contain useful complementary information about both the ionization state and the density field, though this information is secondary to what auto-power spectra provide for these particular parameters. The modest improvements seen here, however, should not be interpreted as evidence that cross-power measurements are of limited scientific value. As we demonstrate in subsequent sections, the information concerning cross-power spectra becomes dramatically more important when considering parameters that are poorly constrained by auto-power shape alone, particularly those related to the properties of ionizing sources. The hierarchy observed here (both better than "auto", which is much better than "cross") reflects the fact that global neutral fraction and density are most directly encoded in the 21cm signal itself, while cross-correlations shine when probing the spatial relationships between ionization and structure formation.

4.2. Impact of galaxy survey parameters

Having established baseline inference performance for our fiducial survey configuration, we now investigate how variations in galaxy-survey design affect the constraining power of 21cm galaxy cross-power measurements. This analysis addresses a practical question faced by the community as to which observational parameters most critically determine scientific return, and which design choices offer the greatest leverage for optimizing future surveys. Understanding these dependencies is essential for guiding observational strategies and evaluating trade-offs among survey area, depth, and redshift precision. We kept the 21cm observational component fixed throughout this analysis, assuming 100h of SKA-Low observations using the AA* array configuration with the moderate foreground avoidance model, as already detailed in Section 4.1. We then systematically varied three main galaxy survey parameters, each spanning a physically motivated range that encompasses both conservative assumptions and optimistic future capabilities. The FOV spans {5, 10, 100} deg², representing various wide-area surveys. The redshift precision varies as $\sigma_z \in \{0.001, 0.01, 0.1\}$, corresponding to spectroscopy, grism, and photometric redshifts. The minimum detectable halo mass threshold takes values of $M_{\text{h,min}} \in \{10^{10}, 10^{10.5}, 10^{11}\} M_{\odot}$, representing extremely deep surveys, deep surveys, and standard galaxy surveys, respectively. These parameter ranges, summarized in Table 2, encompass several forthcoming high-redshift galaxy surveys. Spectroscopic configurations ($\sigma_z = 0.001$) correspond to facilities such as the PFS on Subaru, which can achieve intermediate FOVs (~ 10 deg²) with excellent redshift precision, the proposed WST, which would enable wide-area spectroscopic surveys approaching 100 deg² and the MOON-

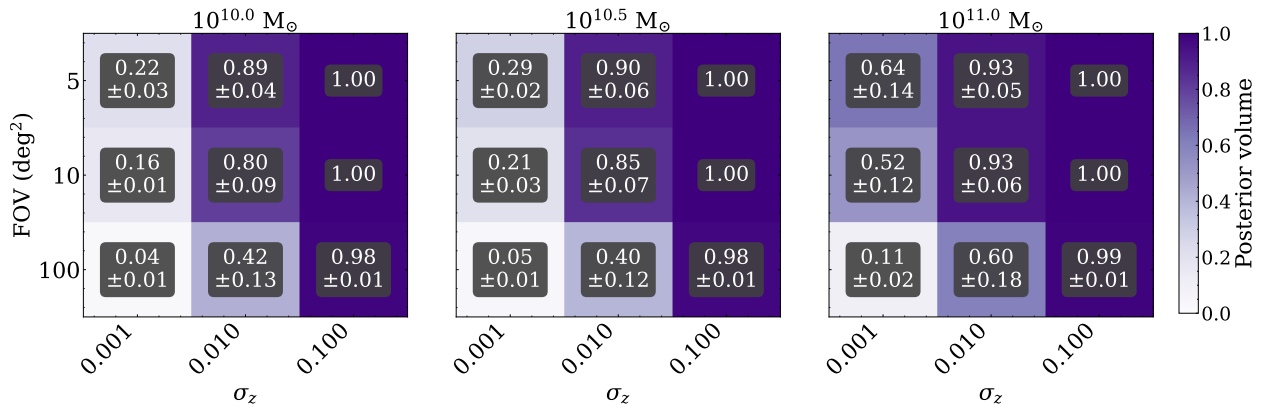


Fig. 4: Posterior volume (averaged over neutral fraction, density, redshifts, and test observations) as a function of galaxy survey parameters assuming moderate 21cm foreground avoidance. Each panel shows results for a different minimum detectable halo mass threshold. The color-scale ranges from white (posterior volume near zero, highly constraining) to dark purple (posterior volume near unity, uninformative), with numerical values annotated in each cell. The left panel shows $M_{h,\min} = 10^{10} M_\odot$, the center panel $10^{10.5} M_\odot$, and the right panel $10^{11} M_\odot$. The FOV increases from top to bottom, and redshift precision improves from right to left. Each cell shows the mean and standard deviations of the PV computed over six independent network training runs; the epistemic uncertainty is discussed in Appendix B.

Table 2: Galaxy survey parameters explored in this section.

Parameter	Values
FOV [deg^2]	[5, 10, 100]
σ_z	[0.001, 0.01, 0.1]
$M_{h,\min} [M_\odot]$	[10^{10} , $10^{10.5}$, 10^{11}]

Notes. All $3^3 = 27$ combinations were evaluated while keeping 21cm observational parameters fixed at the fiducial configuration (100h SKA AA* with moderate foreground avoidance).

RISE program providing spectroscopic observations over $\sim 1 \text{ deg}^2$. Grism observations ($\sigma_z = 0.01$) are within reach of the Nancy Grace Roman Space Telescope, which is expected to survey 5–10 deg^2 in grism mode. Photometric-like precision ($\sigma_z = 0.1$) represents either multiband narrow-band surveys or prism observations. Our fiducial configuration (100 deg^2 , $\sigma_z = 0.001$, $M_{h,\min} = 10^{11} M_\odot$) represents an optimistic but feasible large-area spectroscopic program comparable to an extended PFS campaign or future WST observations, while more conservative, near-term surveys occupy the intermediate parameter space with smaller FOVs and a potentially coarser redshift precision.

Evaluating all $3^3 = 27$ parameter combinations requires training separate EoRFlow models for each survey configuration, as the noise properties and signal characteristics vary substantially across this parameter space. Thanks to the computational efficiency of our framework, this comprehensive survey is feasible. Training a single model requires up to 30 minutes on a single Nvidia A100 GPU depending on survey parameters. To additionally quantify epistemic uncertainty from network training stochasticity, each configuration was trained six times with different random initializations, amounting to 162 models in total. The total computational cost amounts to approximately 70 hours on one GPU, demonstrating the scalability of SBI for survey-optimization studies that would be prohibitive with traditional inference approaches. After training, we evaluated each model on its respective held-out test set of 1000 observations to ensure fair comparison. For each survey configuration, we computed

the PV metric defined in Section 4.1, averaged over both $x_{\text{HI}}(z)$ and $\langle 1 + \delta_{\text{HI}} \rangle(z)$ across all redshift slices and all test observations. This single number summarizes the overall constraining power of each survey design, with values near zero indicating tight constraints and values near unity indicating uninformative observations. Fig. 4 presents these results as a three-panel heat map, with each panel corresponding to a fixed halo mass threshold and showing PV as a function of FOV and redshift precision.

Several clear trends emerge from this parameter exploration. First and most strikingly, redshift precision dominates the constraining power across all configurations. Surveys with poor redshift accuracy ($\sigma_z = 0.1$, corresponding to photometric estimates) yield a PV near unity regardless of FOV or depth, indicating they provide essentially no useful constraints on reionization parameters. Even grism quality ($\sigma_z = 0.01$) shows substantially degraded performance compared to spectroscopy ($\sigma_z = 0.001$). This strong dependence arises because the constraining power of the cross-power spectrum relies on accurate line-of-sight mode measurements. Redshift uncertainties suppress the detection of the line-of-sight component of all Fourier modes, with the effect being most severe for modes corresponding to spatial scales smaller than those probed by σ_z .

Furthermore, the interplay between FOV and halo-mass threshold reveals important survey design considerations. For the most conservative halo-mass limit ($M_{h,\min} = 10^{11} M_\odot$, right panel), only the widest survey (100 deg^2) combined with excellent redshift precision ($\sigma_z = 0.001$) achieves good constraints (PV ~ 0.1). This configuration corresponds to our fiducial WST-type survey from Section 4.1. All other parameter combinations at this halo-mass threshold yield a PV of ≥ 0.5 . While configurations with $\sigma_z = 0.01$ and a large FOV can reduce the prior volume by approximately 50%, this remains modest compared to the spectroscopic case. The physical interpretation is straightforward. At $M_{h,\min} = 10^{11} M_\odot$, galaxy number densities are relatively low, requiring wide areas to accumulate sufficient statistics for measuring the cross-power signal above noise.

Moving to deeper surveys dramatically relaxes the FOV requirement. At $M_{h,\min} = 10^{10.5} M_\odot$ (center panel), even the 10 deg^2 field achieves a PV of around 0.2 when combined with $\sigma_z = 0.001$, representing good constraints. At the most opti-

mistic depth of $M_{h,\min} = 10^{10} M_{\odot}$ (left panel), all field sizes yield excellent constraints ($PV \lesssim 0.2$) provided redshift precision remains high. This trend reflects the rapid increase in galaxy number density toward lower mass thresholds, which enhances the statistical precision of the cross-power measurement through increased mode sampling. Deep surveys thus offer a path to high-quality science with more modest survey areas, though achieving such depths remains observationally challenging.

Intermediate redshift precision shows more modest gains. The central column of each panel shows that $\sigma_z = 0.01$ configurations provide intermediate constraints, though these are substantially degraded compared to spectroscopic precision ($\sigma_z = 0.001$). We caution, however, that these intermediate configurations carry the largest epistemic uncertainty from network training stochasticity (see Appendix B), and their PV values should be interpreted qualitatively rather than as precise quantitative measurements. The overall finding is nonetheless clear: survey strategies should prioritize redshift precision over area or depth when trade-offs are necessary. Investing in high-resolution spectroscopy ($\sigma_z \sim 0.001$) is essential for extracting meaningful constraints from cross-power measurements. Once sufficient galaxy number density is achieved, further improvements from expanding survey area or lowering mass thresholds provide diminishing returns compared to maintaining redshift accuracy.

These results carry important implications for planning future high-redshift galaxy surveys intended for cross-correlation studies with 21cm experiments. Our fiducial configuration (100 deg², $\sigma_z = 0.001$, $M_{h,\min} = 10^{11} M_{\odot}$) represents an ambitious but realistic target for wide spectroscopic surveys that would be possible with the WST. However, the analysis demonstrates that substantial scientific gains are possible with less extreme survey areas if deeper observations can reach $M_{h,\min} \sim 10^{10.5} M_{\odot}$ while maintaining excellent redshift precision. Conversely, photometric or grism surveys will struggle to provide competitive constraints regardless of area or depth, highlighting the critical importance of redshift accuracy for this science case.

Finally, we emphasize that these conclusions are specific to the inference targets considered here, namely the neutral fraction and mean density evolution. As demonstrated in Section 4.4, different astrophysical parameters may exhibit different sensitivities to survey design, with source properties requiring deeper observations. The survey optimization landscape thus depends critically on the scientific questions being addressed, and comprehensive survey design must consider the full range of parameters of interest rather than optimizing for a single metric.

4.3. Information gained from cross-correlation measurements

While the previous sections examined the constraining power of 21cm auto-power and 21cm galaxy cross-power spectra, it is valuable to quantify the per-bin information gained by including the cross-correlation measurements. Specifically, given that 21cm observations provide constraints in each redshift, z , and wave-number, k , bin, we wanted to know how much additional information the cross-power spectrum provides for each (z, k) pair. To address this question, we employed an information-theoretical approach based on mutual information (MI; Cover & Thomas 2006). MI quantifies the reduction in uncertainty about one random variable when another is observed. For discrete random variables X and Y , it is defined as

$$I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \quad (5)$$

where $p(x, y)$ is the joint probability distribution and $p(x)$, $p(y)$ are the marginal distributions. Conditional MI $I(X; Y|Z)$ measures the information gained about X from observing Y when Z is already known, and it can be expressed as

$$I(X; Y|Z) = I(X; [Y, Z]) - I(X; Z), \quad (6)$$

where $I(X; [Y, Z])$ is the MI between X and the joint observation of both Y and Z .

We computed the conditional MI for each parameter, θ_j (where j indexes either the neutral fraction, $x_{\text{HI}}(z)$, or the mean density contrast, $\langle 1 + \delta_{\text{HI}} \rangle(z)$), with respect to the cross-power spectrum conditioned on the auto-power spectrum for each (z, k) bin:

$$I(\theta_j; \text{cross}_{z,k} | \text{auto}_{z,k}) = I(\theta_j; [\text{auto}_{z,k}, \text{cross}_{z,k}]) - I(\theta_j; \text{auto}_{z,k}). \quad (7)$$

This quantity represents the additional information about the parameter θ_j provided by the cross-power at the bin (z, k) beyond what the auto-power already provides. We further computed the fractional information gained as the ratio $I(\theta_j; \text{cross}_{z,k} | \text{auto}_{z,k}) / I(\theta_j; \text{auto}_{z,k})$, which quantifies the relative importance of the cross-power measurement.

We estimated these quantities directly from the simulation dataset. For the fiducial survey configuration (100 deg², $\sigma_z = 0.001$, $M_{h,\min} = 10^{11} M_{\odot}$), we took 1000 test simulations and extracted pairs of (θ, x) , true parameter values, and their corresponding normalized power-spectrum features, using the normalization statistics stored in the trained flow model to ensure consistent preprocessing. For each (z, k) bin, we estimated the marginal MI values $I(\theta_j; \text{auto}_{z,k})$ and $I(\theta_j; \text{cross}_{z,k})$, as well as the joint MI $I(\theta_j; [\text{auto}_{z,k}, \text{cross}_{z,k}])$, using a nonparametric discrete estimator. Both θ_j and each spectral feature were independently discretized into equiprobable quantile bins (32 bins for parameters and 15 bins for spectral features). The marginal MI was then computed from the empirical joint frequency table using Eq. (5). For the joint MI, the pair $(\text{auto}_{z,k}, \text{cross}_{z,k})$ was treated as a single compound discrete variable by encoding the two bin indices as a single index, making no assumption about independence between auto and cross features in the same (z, k) bin. The conditional MI then follows from Eq. (6) by subtraction. This estimator is entirely independent of the posterior approximation of the flow. The per-bin analysis treated different (z, k) bins as being independent of one another; correlations between different bins were not captured. The conditional MI at a given bin therefore reflects the information of that bin marginally, not accounting for any redundancy or synergy with other bins.

Figure 5 shows the fractional MI gain for $x_{\text{HI}}(z)$ (upper panel) and $\langle 1 + \delta_{\text{HI}} \rangle(z)$ (lower panel) as a function of redshift and scale. The color map indicates regions where the cross-power spectrum provides substantial additional information (purple) versus regions where the auto-power spectrum alone is sufficient (white).

Several key trends emerge from this analysis. First, the cross-power spectrum provides a significant conditional information gain across nearly all (z, k) bins, with approximately 78% of bins showing fractional gains exceeding 100% for both parameters. The mean fractional gain is $\sim 270\%$ for x_{HI} and $\sim 280\%$ for $\langle 1 + \delta_{\text{HI}} \rangle$. This seemingly counter-intuitive result must be interpreted carefully. It does not imply that cross-power is more informative than auto-power in absolute terms. Section 4.1 demonstrated that 21cm auto-power alone provides the majority of constraints. Rather, these large fractional values reflect that cross-power measurements provide complementary information breaking specific degeneracies left unresolved by auto-power. While the marginal MI $I(\theta; \text{cross})$ is relatively modest, the conditional MI $I(\theta; \text{cross} | \text{auto})$ can be substantial. Once the 21cm

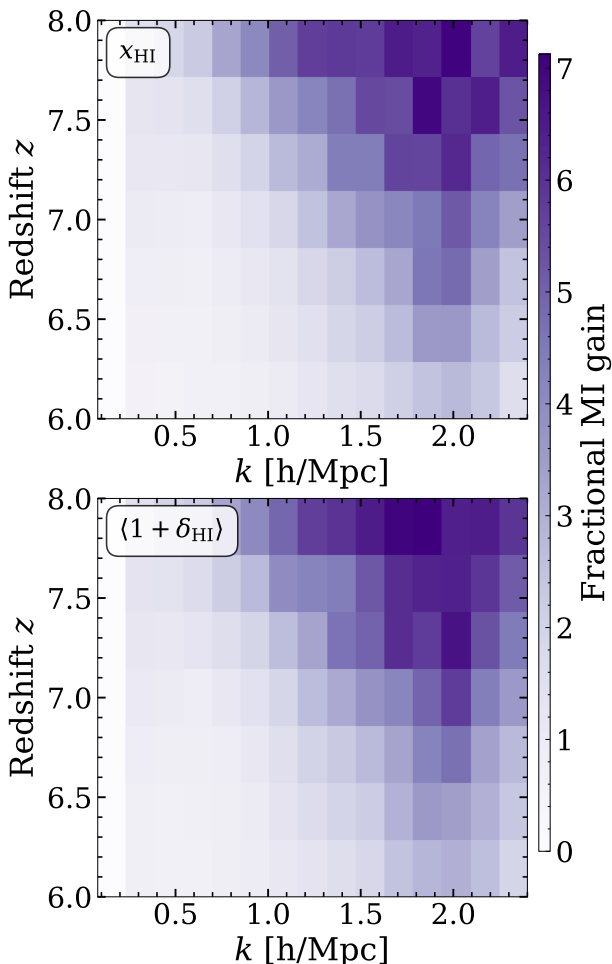


Fig. 5: Fractional mutual information gain from including 21cm galaxy cross-power spectrum measurements in addition to 21cm auto-power ones, as a function of redshift and wave number. Upper panel: Neutral fraction, $x_{\text{HI}}(z)$. Lower panel: Mean density contrast, $\langle 1 + \delta_{\text{HI}} \rangle(z)$. The color-scale indicates the ratio, $I(\theta; \text{cross}_{z,k} | \text{auto}_{z,k}) / I(\theta; \text{auto}_{z,k})$, quantifying the relative additional information provided by cross-correlations. Purple regions indicate high information gain (cross-power provides substantial additional constraints), while white regions show minimal gain (auto-power is sufficient). The cross-power provides significant information across most scales and redshifts, with 78% of bins showing fractional gains exceeding 100%. The information gain increases with redshift and peaks on small scales $k \sim 1 - 2 \text{ Mpc}^{-1}$, reflecting the evolving correlation between ionization morphology and galaxy clustering during reionization.

auto-power establishes the mean ionization state and overall neutral hydrogen distribution, the cross-correlation reveals the spatial relationship between ionization morphology and galaxy clustering; this information is particularly valuable for constraining astrophysical source properties and ionization morphology, which are only weakly determined by power-spectrum shape alone.

Second, there is a clear redshift dependence, with higher fractional gains toward earlier cosmic times (higher z). The fractional gain (averaged over k) increases monotonically from $\sim 135\%$ at $z = 6.0$ to $\sim 450\%$ at $z = 8.0$ for both parameters. This trend reflects the evolving character of reionization: at later times ($z \sim 6$), when reionization is well-advanced, the 21cm

large-scale power and global ionization state provide strong constraints on the remaining parameters. At earlier times ($z \sim 8$) during the onset of reionization, the relationship between the first ionizing sources and the emerging ionized bubbles becomes increasingly critical, and the ability of cross-power to probe this connection yields proportionally larger information gains relative to what auto-power constrains.

Third, the scale dependence shows that large scales ($k \lesssim 0.1 \text{ Mpc}^{-1}$) provide minimal information gains, with fractional MI values near zero. On these scales, cosmic variance dominates both the 21cm and galaxy signals, and their correlation adds little beyond what each observable independently constrains. The information gain increases systematically toward smaller scales, reaching $\sim 290\%$ (averaged over z) at $k \sim 1.2 \text{ Mpc}^{-1}$ and $\sim 370\%$ on the smallest scales probed ($k \sim 2.4 \text{ Mpc}^{-1}$). These intermediate to small scales are where ionization morphology, bubble sizes, shapes, and their correlation with the underlying density field traced by galaxies, are most directly manifested. The cross-power on these scales provides crucial constraints on how ionized regions grow around galaxies; this information complements the auto-power global characterization of the neutral fraction distribution. Comparing the two parameters, $\langle 1 + \delta_{\text{HI}} \rangle$ shows slightly higher fractional gains than x_{HI} (by a factor of ~ 1.03), though the difference is modest. The mean conditional MI values are 1.28 nats for x_{HI} and 1.30 nats for $\langle 1 + \delta_{\text{HI}} \rangle$, indicating comparable absolute information content from cross-power for both quantities.

These substantial conditional MI gains translate to modest but meaningful improvements in posterior precision for global reionization parameters, where auto-power already provides strong baseline constraints. The large conditional MI values on intermediate to small scales and at higher redshifts—precisely where ionization morphology and source-bubble correlations are most important—suggest that cross-power measurements carry crucial complementary information for breaking degeneracies unresolved by auto-power alone. This pattern is directly supported by gradient-based saliency analysis of the trained neural network (Appendix C), which reveals that the 21cm auto-power network focuses attention on low redshifts across all scales, while the cross-power network distributes attention across all redshifts but concentrates on intermediate scales. Since the 21cm auto-power already captures low-redshift information efficiently, the cross-power naturally provides the greatest information gain at higher redshifts where auto-power constraints are weaker, explaining why the conditional MI peaks precisely in these regimes. In the following section, we extend our analysis beyond the neutral fraction and density evolution to examine how well cross-power measurements constrain the astrophysical properties of the ionizing sources themselves.

4.4. Inferring reionization source properties

To demonstrate the unique potential of 21cm galaxy cross-correlations for constraining source properties, we retrained our framework to infer the four astrophysical parameters, $\{f_{\text{esc},10}, f_{*,10}, \alpha_{\text{esc}}, \alpha_*\}$ (Eq.(2)), from the power-spectrum measurements. Here $f_{\text{esc},10}$ and $f_{*,10}$ denote the escape fraction and star formation efficiency normalized to halos with masses of $10^{10} M_{\odot}$, while the power-law exponents describe their mass dependence. These parameters directly control where and how efficiently ionized bubbles form around galaxies, making them natural targets for cross-power inference but challenging to extract from 21cm auto-power alone, which primarily traces the integrated ionization state rather than source-bubble correlations.

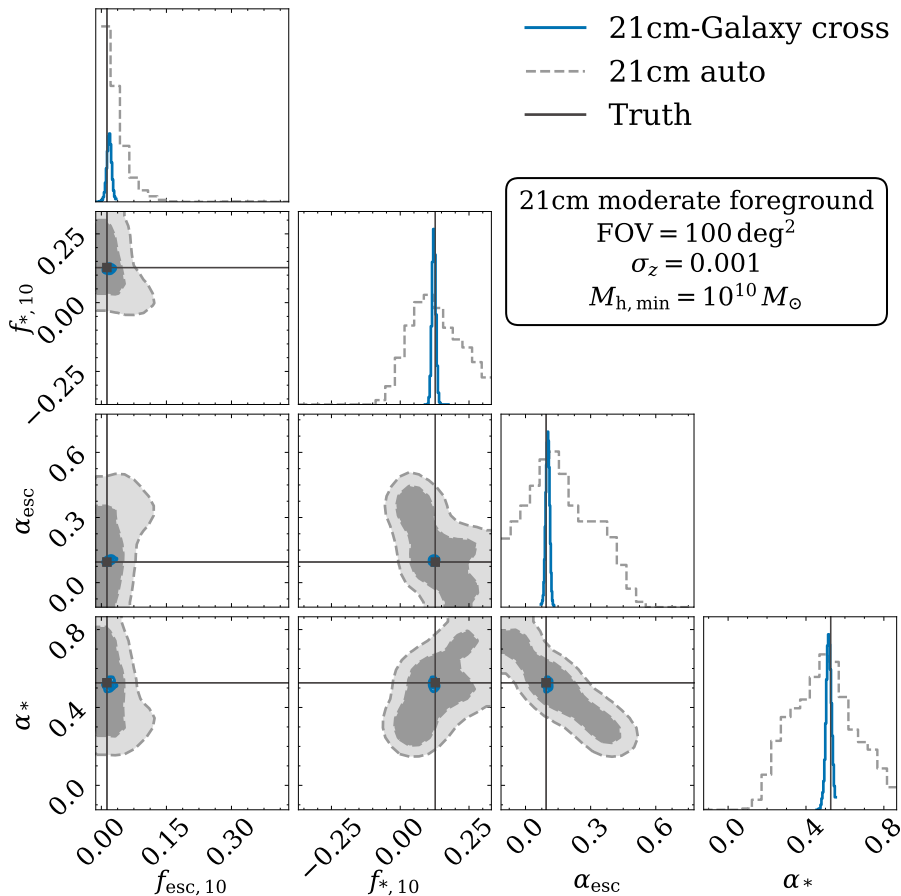


Fig. 6 presents the inference results as an example corner plot using the fiducial survey configuration (100 deg^2 , $\sigma_z = 0.001$) but with an optimistic halo mass threshold of $M_{h,\text{min}} = 10^{10} M_\odot$. This choice probes the deepest future galaxy surveys and serves as a proof of concept for the method’s potential. The 21cm auto-power spectrum (dashed gray contours) shows no constraining power beyond revealing strong parameter degeneracies, particularly between α_* and α_{esc} where different ionization morphologies yield identical EoR timelines. Across the whole test set the R^2 scores (where 1 indicates perfect inference accuracy) range from 0.08 to 0.47, and the relative PV is between 60% and 87%. The posteriors remain broad and weakly correlated with the true values, indicating that the auto-power spectrum alone, while highly informative of the neutral fraction evolution, carries insufficient information to break degeneracies among source parameters that produce similar global ionization histories through different combinations of escape fractions and star formation efficiencies.

In stark contrast, the 21cm galaxy cross-power spectrum (solid blue contours) provides excellent constraints on all four parameters. Tight, circular contours indicate precise and non-degenerate constraints, accurately recovering the true parameters. Across the test dataset the R^2 scores consistently exceed 0.92, and the mean relative PV is $\sim 11\%$ across all quantities. This dramatic improvement can be understood physically. While the auto-power measures the spatial distribution of neutral hydrogen, the cross-power directly probes the correlation between ionized regions and galaxy positions. This correlation is highly sensitive to how ionizing photons escape from galaxies (through f_{esc}) and how efficiently stars form in halos of different masses (through f_*). Galaxies with higher escape fractions cre-

Fig. 6. Inference of astrophysical source parameters from 21cm auto-power and 21cm galaxy cross-power spectra for a randomly chosen sample assuming moderate 21cm foreground avoidance. Corner plots show the marginalized posteriors for the four parameters controlling escape fraction and star formation efficiency: $\log_{10} f_{\text{esc},10}$, $\log_{10} f_{*,10}$, α_{esc} , and α_* . Solid blue contours show cross-power constraints ($R^2 > 0.92$, mean posterior volume of $\sim 11\%$ across the entire test dataset). Dashed gray contours show auto-power constraints, which are virtually uninformative ($R^2 < 0.47$, posterior volume of $> 60\%$ across the entire test dataset). Results assume an optimistic future survey with a FOV of 100 deg^2 , spectroscopic redshift uncertainty of $\sigma_z = 0.001$, and a halo mass detection threshold of $M_{h,\text{min}} = 10^{10} M_\odot$ (for comparison with optimistic foreground removal at $M_{h,\text{min}} = 10^{11} M_\odot$, see Fig. D.1). The dramatic contrast demonstrates that cross-correlations provide access to source properties fundamentally inaccessible to 21cm auto-power alone, though this requires detecting faint galaxies below currently planned survey sensitivities.

ate larger ionized bubbles around themselves, strengthening the anti-correlation signal in the cross-power, while the mass dependencies determine which galaxy populations contribute most to reionization and thus shape the scale dependence of the cross-power.

Combining both observables (cross+auto) yields performance levels comparable to that of cross-power alone on average, though with substantially larger run-to-run variance ($\text{PV} = 0.24 \pm 0.12$ compared to 0.11 ± 0.02 for cross-power alone). This increased epistemic noise likely reflects the challenge of training a single neural density estimator on two observables with vastly different information; the cross-power is highly informative of source properties, while the auto-power is not, and the network must allocate its capacity to model the joint likelihood of both. The large variance suggests that the network sometimes successfully focuses on the informative cross-power features and sometimes does not, depending on initialization. A dedicated architecture optimized for the combined input, or an adaptive weighting scheme, would likely yield more consistent results. We quantified the epistemic uncertainty by retraining each model across multiple random initializations. The 21cm auto-power models show negligible run-to-run variance ($\text{PV} = 0.76 \pm 0.01$), consistent with all initializations converging to the same uninformative posterior. The cross-power models also show small variance ($\text{PV} = 0.11 \pm 0.02$), with all qualitative conclusions remaining robust across runs. The epistemic uncertainty across all configurations is discussed further in Appendix B.

The results presented here assume an optimistic $M_{h,\text{min}} = 10^{10} M_\odot$ halo mass threshold, corresponding to extremely deep future galaxy surveys that can detect faint star-forming galaxies down to this limit across 100 deg^2 . Increasing the threshold

to $M_{h,\min} = 10^{10.5} M_{\odot}$ yields only moderately degraded constraints ($PV = 0.17 \pm 0.03$), suggesting that surveys reaching this intermediate depth already provide meaningful source-property constraints. At the more commonly assumed $M_{h,\min} = 10^{11} M_{\odot}$ threshold however, the source parameter constraints become marginal ($PV = 0.50 \pm 0.10$). On the contrary, Appendix D demonstrates that optimistic 21cm foreground removal strategies, which recover wedge modes, can restore tight constraints on source properties ($PV = 0.18 \pm 0.04$) even at the $M_{h,\min} = 10^{11} M_{\odot}$ threshold, offering a complementary pathway to achieving source-property constraints without requiring detection of the faintest galaxies. We note that additional uncertainties arise when detecting galaxies via, for example, Lyman- α emission, including the stochasticity of luminosities at fixed halo mass and variations in IGM transmission. These effects contribute to the effective shot noise and scatter in galaxy bias. However, for typical survey sensitivities, such contributions remain subdominant compared to uncertainties from survey completeness, set by the halo mass (or luminosity) detection threshold, and 21cm signal foreground contamination.

The contrast between Fig. 5 (showing large conditional MI for neutral fraction and density) and Fig. 6 (showing weak auto-power constraints on source properties) illustrates a key distinction. While cross-power provides complementary information that modestly improves constraints on parameters already well measured by auto-power, it provides essential information for parameters that are degenerate or invisible to auto-power alone. This dual role of enhancement and enablement positions 21cm galaxy cross-correlations as a uniquely powerful probe of cosmic reionization. By simultaneously confirming the cosmological 21cm signal, providing robustness against foreground contamination, and accessing source properties inaccessible to 21cm measurements alone, cross-power studies represent not merely an incremental improvement over single-probe analyses, but a qualitatively new observational window into the EoR.

5. Conclusions

In this work, we developed an SBI framework to quantify the constraining power of 21cm galaxy cross-correlations during the EoR. Our methodology combines realistic forward modeling of noisy, 21cm galaxy cross-power spectra generated with 21cmFASTv4 and incorporates instrumental and survey effects following Hutter & Heneka (2026), with the EoRFlow SBI framework of Pietschke et al. (2025). This approach enables likelihood-free inference on astrophysical parameters governing reionization without requiring analytic approximations to the intractable data likelihood. The inference network operates directly on spherically averaged 1D power spectra using conditional normalizing flows with rational quadratic spline coupling layers, providing increased expressivity compared to earlier implementations. Our main findings are as follows.

Inference of global reionization properties: Assuming 100h of SKA AA* observations with moderate foreground avoidance and a WST-type fiducial galaxy survey ($FOV = 100\text{deg}^2$, $\sigma_z = 0.001$, $M_{h,\min} = 10^{11} M_{\odot}$), we demonstrate that 21cm galaxy cross-power spectra alone yield unbiased constraints on the neutral hydrogen fraction $x_{\text{HI}}(z)$ and the mean overdensity in neutral regions $\langle 1 + \delta_{\text{HI}} \rangle(z)$ across $z = 6-8$. While these constraints are weaker than those obtained from the 21cm auto-power spectrum, they provide a physically independent probe of the reionization history. When combined with auto-power measurements, poste-

rior volumes decrease by approximately 20–30% compared to auto-power ones alone, confirming that cross-correlations contribute genuinely complementary information rather than redundant constraints.

Galaxy-survey requirements: Our systematic exploration of survey parameter space reveals that redshift precision is the dominant factor determining the scientific utility of cross-correlation measurements. Spectroscopic-quality redshifts ($\sigma_z \sim 0.001$) provide optimal constraints. Intermediate precision achievable with grisms ($\sigma_z \sim 0.01$) can yield meaningful constraints when combined with large survey areas, though these configurations carry the largest epistemic uncertainty and their PV values should be interpreted qualitatively. In all cases they fall short of the factor of ~ 10 improvement achievable with full spectroscopy. Photometric uncertainties ($\sigma_z \sim 0.1$) severely degrade constraints regardless of survey area or depth. FOV and minimum detectable halo mass play secondary but important roles, with deeper surveys ($M_{h,\min} \lesssim 10^{10.5} M_{\odot}$) partially compensating for smaller survey areas. These findings provide concrete guidance for designing future galaxy surveys intended for cross-correlation studies with 21cm experiments.

Complementary information from cross-correlations: Using mutual information analysis, we quantified where cross-power measurements provide the greatest information gain beyond what the 21cm auto-power spectrum constrains. The additional information is concentrated at intermediate to small scales ($k \sim 0.5-2\text{Mpc}^{-1}$), corresponding to characteristic ionized bubble sizes, and increases toward higher redshifts during the early phases of reionization. This pattern reflects the sensitivity of cross-correlations to the evolving spatial relationship between ionizing sources and the surrounding ionization morphology.

Constraints on ionizing source properties: The most significant finding is that 21cm galaxy cross-correlations provide access to astrophysical source properties that are fundamentally inaccessible to 21cm auto-power measurements alone. We showed that the escape fraction ($f_{\text{esc},10}$, α_{esc}) as well as the star formation efficiency parameters ($f_{*,10}$, α_*) can be tightly constrained from cross-power spectra, achieving $R^2 > 0.92$ and mean posterior volumes of $\sim 11\%$. In contrast, the 21cm auto-power spectrum yields $R^2 < 0.5$ and posterior volumes exceeding 60%, owing to strong degeneracies among source models that produce similar global ionization histories. Notably, this unique constraining power of 21cm galaxy cross-correlations can be achieved through either deep surveys detecting faint galaxies down to $M_{h,\min} \sim 10^{10} M_{\odot}$ under moderate foreground treatment, or more conservative mass thresholds ($M_{h,\min} \sim 10^{11} M_{\odot}$) when paired with optimistic 21cm foreground removal (Appendix D). This capability transforms cross-correlations into a powerful diagnostic of the reionization source population.

Implications for observational programs: Our results carry direct implications for the planning and prioritization of forthcoming surveys. For facilities such as the WST and the PFS, spectroscopic observations over areas of $\gtrsim 10-100\text{deg}^2$ with high redshift precision will be essential to exploit the full potential of 21cm galaxy cross-correlations. Critically, tight constraints on astrophysical source properties can be achieved through either deep surveys reaching $M_{h,\min} \sim 10^{10}-10^{10.5} M_{\odot}$ under mod-

erate 21cm foreground treatment, or surveys targeting brighter galaxies ($M_{h,\min} \sim 10^{11} M_{\odot}$) paired with optimistic foreground removal strategies (Appendix D). This flexibility allows observational programs to optimize either galaxy survey sensitivity or 21cm foreground cleaning capabilities depending on technological developments and resource allocation.

Future directions: This work establishes a foundation for inference on 21cm galaxy cross-correlations, but several extensions promise further scientific gains. Our analysis assumed moderate 21cm foreground avoidance. Optimistic foreground removal strategies could substantially enhance constraining power. As demonstrated in Appendix D, this relaxation enables tight constraints on astrophysical source properties even when galaxy surveys reach only $M_{h,\min} \sim 10^{11} M_{\odot}$. These findings align with the analysis of Hutter & Heneka (2026), which showed that optimistic foreground scenarios dramatically relax survey requirements for detecting the cross-correlation signal itself. The 1D power spectrum employed here, while providing a favorable S/N, discards information encoded in the anisotropic and non-Gaussian structure of the cross-correlation signal. Moving beyond power spectra to leverage topological summary statistics, such as Minkowski functionals or Betti numbers, could capture the morphological information encoded in the shapes and connectivity of ionized regions. Machine-learning approaches offer a natural pathway to extract such features, either through learned summary networks that compress the full 3D cross-correlation field or through neural-network architectures designed to operate directly on spatial data.

In summary, 21cm galaxy cross-correlations occupy a unique position in the landscape of reionization probes. They provide an independent confirmation of the cosmological 21cm signal, are robust against foreground contamination, and unlock constraints on source properties that are inaccessible to 21cm auto-power measurements alone. As both 21cm experiments and high-redshift galaxy surveys approach the sensitivities required for cross-correlation detections, the SBI framework developed here offers a flexible and scalable approach to extracting maximal astrophysical insight from this powerful multi-tracer view of the EoR.

Data availability

The code used to compute the 21cm galaxy cross-power spectrum uncertainties is publicly available at github.com/annehutter/21cm_gal_uncertainties. This work builds upon the EoRFlow framework (Pietschke et al. 2025), available at github.com/astro-ML/EoRFlow. The project-specific analysis scripts and trained neural-network models can be found at github.com/astro-ML/21cmgalaxy-SBI.

Acknowledgements. We would like to thank Maike Voelkel, Vrund Patel and M. Paola Vaccaro for helpful discussions. We also thank the anonymous referee for their insightful comments, which improved the clarity and robustness of this work. YP's and CH's work is funded by the Volkswagen Foundation. This work was supported by the DFG under Germany's Excellence Strategy EXC 2181/1 - 390900948 The Heidelberg STRUCTURES Excellence Cluster. The authors acknowledge support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 37/935-1 FUGG.

References

Abdurashidova, T. H. C. Z., Adams, T., Aguirre, J. E., et al. 2023, *ApJ*, 945, 124

- Abdurashidova, Z., Adams, T., Aguirre, J. E., et al. 2026, *ApJ*, 998, 33
- Abdurashidova, Z., Aguirre, J. E., Alexander, P., et al. 2022, *ApJ*, 924, 51
- Abdurashidova, Z., Aguirre, J. E., Alexander, P., et al. 2022, *ApJ*, 925, 221
- Acharya, A., Mertens, F., Ciardi, B., et al. 2024, *MNRAS*, 534, L30
- Aghanim, N. et al. 2020, *A&A*, 641, A6, [Erratum: *A&A* 652, C4 (2021)]
- Alsing, J., Charnock, T., Feeney, S., & Wandelt, B. 2019, *MNRAS*, 488, 4440
- Ansel, J., Yang, E., He, H., et al. 2024, in *Proceedings of ASPLOS (ACM)*
- Beane, A., Villaescusa-Navarro, F., & Lidz, A. 2019, *ApJ*, 874, 133
- Ceccotti, E., Offringa, A. R., Mertens, F. G., et al. 2025, *MNRAS*, 544, 1255
- Cerardi, N., Giri, S. K., Bianco, M., et al. 2025, Submitted to *MNRAS*, arXiv:2511.11568
- Cole, A., Miller, B. K., Witte, S. J., et al. 2022, *JCAP*, 2022, 004
- Cooper, N., Norregaard, C., Meriot, R., & Pritchard, J. R. 2026, *MNRAS*, 548, stag577
- Cover, T. M. & Thomas, J. A. 2006, *Elements of Information Theory*, 2nd edn. (Hoboken, NJ: Wiley-Interscience)
- Davies, J. E., Mesinger, A., & Murray, S. G. 2025, *A&A*, 701, A236
- DeBoer, D. R., Parsons, A. R., Aguirre, J. E., et al. 2017, *PASP*, 129, 045001
- Durkan, C., Bekasov, A., Murray, I., & Papamakarios, G. 2019, in *Proceedings of NeurIPS*, Vol. 32 (Curran Associates, Inc.)
- Furlanetto, S. R. & Lidz, A. 2007, *ApJ*, 660, 1030
- Furlanetto, S. R., Oh, S. P., & Briggs, F. H. 2006, *Phys. Rep.*, 433, 181
- Gagnon-Hartman, S., Davies, J. E., & Mesinger, A. 2025, *A&A*, 699, A131
- Greene, J., Bezanson, R., Ouchi, M., Silverman, J., & the PFS Galaxy Evolution Working Group. 2022, *The Prime Focus Spectrograph Galaxy Evolution Survey*
- Heneka, C. & Cooray, A. 2021, *MNRAS*, 506, 1573
- Heneka, C., Cooray, A., & Feng, C. 2017, *ApJ*, 848, 52
- Heneka, C. & Mesinger, A. 2020, *MNRAS*, 496, 581
- HERA Collaboration, Abdurashidova, Z., Adams, T., et al. 2023, *ApJ*, 945, 124
- Hutter, A., Dayal, P., Müller, V., & Trott, C. M. 2017, *ApJ*, 836, 176
- Hutter, A. & Heneka, C. 2026, *A&A*, 707, A286
- Hutter, A., Heneka, C., Dayal, P., et al. 2023a, *MNRAS*, 525, 1664
- Hutter, A., Trebitsch, M., Dayal, P., et al. 2023b, *MNRAS*, 524, 6124
- Kubota, K., Yoshiura, S., Takahashi, K., et al. 2018, *MNRAS*, 479, 2754
- La Plante, P., Mirocha, J., Gorce, A., Lidz, A., & Parsons, A. 2023, *ApJ*, 944, 59
- Lidz, A., Zahn, O., Furlanetto, S. R., et al. 2009, *ApJ*, 690, 252
- Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., & Le, M. 2023, in *Proceedings of ICLR*
- Loshchilov, I. & Hutter, F. 2019, in *Proceedings of ICLR*
- Mainieri, V., Anderson, R. I., Brinchmann, J., et al. 2024, *The Wide-field Spectroscopic Telescope (WST) Science White Paper*
- Maiolino, R., Cirasuolo, M., Afonso, J., et al. 2020, *The Messenger*, 180, 24
- Mertens, F. G., Mevius, M., Koopmans, L. V. E., et al. 2020, *MNRAS*, 493, 1662
- Mesinger, A., Furlanetto, S., & Cen, R. 2011, *MNRAS*, 411, 955
- Moriwaki, K., Beane, A., & Lidz, A. 2024, *MNRAS*, 530, 3183
- Murray, S. G., Greig, B., Mesinger, A., et al. 2020, *JOSS*, 5, 2582
- Ore, A., Heneka, C., & Plehn, T. 2025, *SciPost Phys.*, 18, 155
- Park, J., Kim, H.-S., Wyithe, J. S. B., & Lacey, C. G. 2014, *MNRAS*, 438, 2474
- Park, J., Mesinger, A., Greig, B., & Gillet, N. 2019, *MNRAS*, 484, 933–949
- Parsons, A. R., Backer, D. C., Foster, G. S., et al. 2010, *AJ*, 139, 1468
- Pietschke, Y., Heneka, C., Schlenker, T., Ore, A., & Schosser, B. 2025, *JCAP*, 2025, 039
- Pober, J. C., Liu, A., Dillon, J. S., et al. 2014, *ApJ*, 782, 66
- Pober, J. C., Parsons, A. R., DeBoer, D. R., et al. 2013, *AJ*, 145, 65
- Saxena, A., Meerburg, P. D., Weniger, C., Acedo, E. d. L., & Handley, W. 2024, *RASTI*, 3, 724
- Schosser, B., Heneka, C., & Plehn, T. 2025, *SciPost Phys. Core*, 8, 037
- Schosser, B., Heneka, C., & Schäfer, B. M. 2026, *JCAP*, 2026, 086
- Sobacchi, E., Mesinger, A., & Greig, B. 2016, *MNRAS*, 459, 2741
- Tingay, S. J., Goeke, R., Bowman, J. D., et al. 2013, *PASA*, 30, e007
- Tong, A., Fatras, K., Malkin, N., et al. 2024, in *Proceedings of TMLR*
- Trott, C. M., Jordan, C. H., Midgley, S., et al. 2020, *MNRAS*, 493, 4711
- van Haarlem, M. P., Wise, M. W., Gunst, A. W., et al. 2013, *A&A*, 556, A2
- Villaescusa-Navarro, F., Genel, S., Anglés-Alcázar, D., et al. 2022, *ApJS*, 259, 61
- Vrbanc, D., Ciardi, B., Jelić, V., et al. 2020, *MNRAS*, 492, 4952
- Vrbanc, D., Ciardi, B., Jelić, V., et al. 2016, *MNRAS*, 457, 666
- Wang, Y., Zhai, Z., Alavi, A., et al. 2022, *ApJ*, 928, 1
- Weinberger, L. H., Kulkarni, G., & Haehnel, M. G. 2020, *MNRAS*, 494, 703
- Yoshiura, S., Line, J. L. B., Kubota, K., Hasegawa, K., & Takahashi, K. 2018, *MNRAS*, 479, 2767
- Yoshiura, S., Pindor, B., Line, J. L. B., et al. 2021, *MNRAS*, 505, 4775

Appendix A: Dimensionality of the power spectrum

Throughout this work, we employ spherically averaged (1D) power spectra for parameter inference. While the cylindrically averaged (2D) power spectrum retains directional information along the line-of-sight and perpendicular directions, which in principle encodes additional sensitivity to redshift-space distortions and ionization anisotropies, we find that the enhanced S/N achieved through spherical averaging yields superior inference performance in practice.

Table A.1 presents a quantitative comparison of inference quality using 1D versus 2D power spectra for the fiducial survey configuration (FOV = 100deg², $\sigma_z = 0.001$, $M_{h,\min} = 10^{11}M_\odot$). We evaluate performance using the posterior volume (PV) metric, defined as the standard deviation of posterior samples normalized by the prior standard deviation, computed over 1000 test observations. Lower values indicate tighter constraints.

Table A.1: Comparison of 1D and 2D power spectrum performance for different data modes.

Dim.	Mode	PV (x_{HI})	PV($\langle 1 + \delta_{\text{HI}} \rangle$)
1D	21cm	0.0248 ± 0.0015	0.0677 ± 0.0043
1D	Cross	0.0893 ± 0.0205	0.1223 ± 0.0234
1D	Both	0.0176 ± 0.0016	0.0496 ± 0.0032
2D	21cm	0.0231 ± 0.0080	0.0769 ± 0.0411
2D	Cross	0.1758 ± 0.0206	0.2643 ± 0.0203
2D	Both	0.0193 ± 0.0016	0.0567 ± 0.0054

Notes. Metrics are computed over the 1000 test samples with epistemic network uncertainties. Best values (lowest PV) are highlighted with boxes.

For the 21cm auto-power spectrum, 1D and 2D performance are comparable, with 2D achieving modestly better constraints for x_{HI} (PV = 0.023 ± 0.008 versus 0.025 ± 0.002 for 1D), though the difference is within the epistemic noise. However, for the 21cm galaxy cross-power spectrum, the pattern reverses dramatically. The 2D cross-power yields substantially degraded constraints (PV = 0.176 ± 0.021 for x_{HI} versus 0.089 ± 0.021 for 1D, and 0.264 ± 0.020 versus 0.122 ± 0.023 for $\langle 1 + \delta_{\text{HI}} \rangle$). This behavior arises because the cross-correlation measurement is subject to additional uncertainty sources beyond those affecting the auto-power, and the directional information retained in the 2D format does not compensate for the loss of S/N from reduced mode averaging.

When combining both auto-power and cross-power measurements, the 1D version delivers the tightest constraints (PV = 0.018 ± 0.002 for x_{HI} and 0.050 ± 0.003 for $\langle 1 + \delta_{\text{HI}} \rangle$), marginally outperforming the 2D combination (PV = 0.019 ± 0.002 and 0.057 ± 0.005), though the difference is within the epistemic noise for x_{HI} . The clear advantage of 1D is most pronounced for the cross-power alone, where the S/N gain from spherical averaging outweighs the loss of directional information. These results justify our choice to use 1D power spectra throughout the main analysis.

Appendix B: Epistemic uncertainty

In Section 4.2 we systematically varied the galaxy survey parameters and studied the constraining power of the 21cm galaxy cross-power spectrum by means of the posterior volume (PV). The PV metric carries an intrinsic uncertainty arising from the

stochasticity of neural network training, which we quantify here. For each of the 27 survey configurations, we trained 6 networks with different random initializations, keeping all other hyperparameters fixed, amounting to 162 trained models in total. The same procedure was applied to the three models in Section 4.1, where epistemic noise is small given the fiducial survey configuration, and to the models in Section 4.4, where the combined cross+auto model shows the largest epistemic noise (PV = 0.24 ± 0.12), likely due to the architectural challenges of training on two observables with vastly different information content. The resulting mean PV and standard deviation across runs are shown directly in Fig. 3 and Fig. 4. We highlight that this comprehensive study is only feasible thanks to the computational efficiency of our SBI framework; traditional inference approaches would render such a large-scale exploration prohibitive. Furthermore, we verified that sampling uncertainties arising from the finite test set size are negligible compared to the epistemic noise reported here, confirming that network training stochasticity is the dominant source of uncertainty in the PV metric. The variance in PV follows a clear hierarchy determined by the constraining power of the survey. Well-constrained configurations (e.g. FOV = 100deg², $\sigma_z = 0.001$, $M_{h,\min} = 10^{10}M_\odot$) yield PV = 0.04 ± 0.01: the data are highly informative, the network learns consistently regardless of initialization, and the epistemic noise is low. At the other extreme, configurations with very high observational noise (e.g. $\sigma_z = 0.1$) yield PV ≈ 1.0 with negligible variance: the data are entirely uninformative and all initializations converge to the same uninformative posterior. The largest epistemic noise arises in intermediate configurations (e.g. FOV = 100deg², $\sigma_z = 0.01$, $M_{h,\min} = 10^{11}M_\odot$, PV = 0.60 ± 0.18): here the data contain some information but the learning problem is hard enough that the network only succeeds from favorable initializations, leading to large run-to-run variance without further architecture tuning.

We attribute this behavior to two effects. First, networks trained on uninformative data face a harder optimization problem with loss curves more sensitive to initialization. Second, the network architecture and training hyperparameters were optimized for the fiducial survey configuration; applying them unchanged across all 27 configurations is an intentional design choice to isolate the effect of survey parameters, but it means the training is suboptimal for the most challenging configurations. Overall, Fig. 4 provides a clear qualitative hierarchy of survey performance that is robust across all training runs. However, particularly for intermediate configurations ($\sigma_z = 0.01$) where epistemic noise is largest, the individual PV values should not be interpreted as precise quantitative measurements, and conclusions drawn from these configurations should remain qualitative in nature.

Appendix C: Saliency analysis

To understand which regions of the input power spectra most strongly influence parameter inference by the neural network, we perform a saliency analysis by computing gradients of the log-probability with respect to the input features. This approach reveals where the network focuses its attention across different scales and redshifts, providing insight into which observational modes carry the most relevant information for constraining reionization parameters.

For a trained normalizing flow, the gradient $\nabla_{\mathbf{x}} \log p(\theta|\mathbf{x})$ quantifies the sensitivity of the posterior probability to small changes in each input feature. We compute the average absolute gradient across all test samples, creating saliency maps that

highlight which (k, z) bins in the power spectrum the network weights most heavily during inference.

Fig. C.1 presents saliency maps for inference on $x_{\text{HI}}(z)$ and $\langle 1 + \delta_{\text{HI}} \rangle(z)$ using the 21cm auto-power spectrum (upper panel) and the 21cm galaxy cross-power spectrum (lower panel).

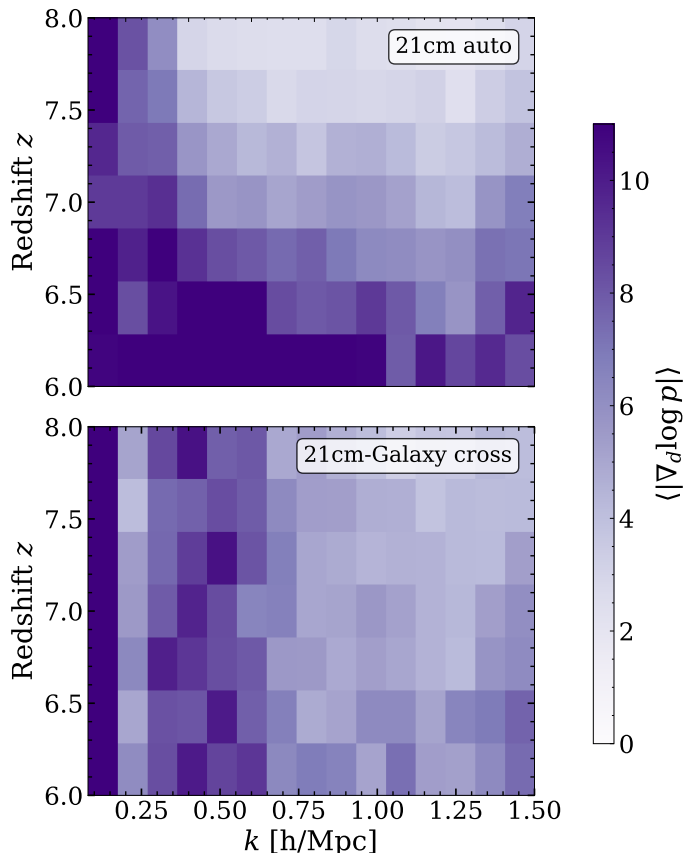


Fig. C.1: Saliency analysis for the inference of $x_{\text{HI}}(z)$ and $\langle 1 + \delta_{\text{HI}} \rangle(z)$ from the 21cm auto-power spectrum (left) and the 21cm galaxy cross-power spectrum (right). The colormap shows the average absolute gradient of the log-probability with respect to the input power spectrum, indicating which redshift bins and spatial scales the network attends to most strongly. Both observables show high attention to the lowest k bin across all redshifts, corresponding to the foreground-masked region that the network must learn to disregard. The 21cm auto-power focuses primarily on low redshifts across all scales, while the cross-power shows more uniform attention across redshifts but concentrates on intermediate scales ($k \sim 0.3\text{-}0.7 h\text{Mpc}^{-1}$), consistent with the characteristic sizes of ionized regions.

Both maps exhibit high gradient values at the lowest k modes across all redshifts, corresponding to the foreground wedge region excluded by our observational mask. The network must learn to recognize and disregard this masked region, which contains no physical information about reionization, explaining the elevated attention in this regime. Beyond the masked region, the two observables display distinct attention patterns. For the 21cm auto-power spectrum, the network concentrates attention on the lowest redshifts ($z \sim 6\text{-}7$) across a broad range of scales. This pattern reflects the fact that late-stage reionization, when the neutral fraction is rapidly declining and large-scale power is maximized, provides the strongest constraints on global reionization properties. In contrast, the 21cm galaxy cross-power spectrum shows relatively uniform attention across all redshifts

but focuses on intermediate scales ($k \sim 0.3\text{-}0.7 h\text{Mpc}^{-1}$). These scales correspond to the characteristic sizes of ionized regions and the transition between regimes where galaxies trace ionized bubbles versus underlying density fluctuations, consistent with the physical expectations discussed in Section 4.3.

These findings align with both the mutual information analysis presented in the main text and physical intuition about reionization signatures. The sensitivity of the cross-power to intermediate scales during all phases of reionization reflects its direct probe of ionization morphology, while the emphasis of the auto-power on late times captures the evolution of global neutral fraction. Nevertheless, we emphasize that these patterns are emergent properties of the trained network architecture and do not constitute a fundamental decomposition of information content. Alternative network designs or summary statistics might access the same underlying information through different pathways.

Appendix D: Optimistic 21cm foregrounds

Section 4.4 demonstrated that 21cm galaxy cross-power spectra provide tight constraints on astrophysical source properties when assuming deep galaxy surveys capable of detecting halos down to $M_{\text{h,min}} = 10^{10} M_{\odot}$ under moderate 21cm foreground avoidance. However, this requires detecting extremely faint galaxies that may lie beyond the reach of currently planned survey facilities. Here we show that optimistic 21cm foreground removal strategies offer an alternative pathway to achieving comparable constraints on source properties even with the more conservative halo mass threshold of $M_{\text{h,min}} = 10^{11} M_{\odot}$.

Under moderate foreground avoidance, we mask contaminated modes in the foreground wedge following the prescription of Pober et al. (2014), which excludes modes below the horizon limit plus a conservative $0.1 h\text{Mpc}^{-1}$ buffer. This approach sacrifices the large-scale modes most contaminated by foreground emission to ensure robust detection of the cosmological signal. However, recent advances in foreground removal techniques (e.g., Mertens et al. 2020; Acharya et al. 2024) suggest that more aggressive cleaning strategies may successfully recover these wedge modes, particularly for cross-correlation measurements where foreground contamination is partially suppressed through decorrelation between 21cm and galaxy fields. To explore this optimistic scenario, we retrain our inference framework assuming all modes down to the primary beam scale can be recovered (the "optimistic foreground model" of 21cmSense). We apply this to the fiducial survey configuration ($\text{FOV} = 100 \text{ deg}^2$, $\sigma_z = 0.001$, $M_{\text{h,min}} = 10^{11} M_{\odot}$) and perform inference on the four astrophysical source parameters $\{f_{\text{esc},10}, f_{*,10}, \alpha_{\text{esc}}, \alpha_{*}\}$.

Fig. D.1 presents the inference results, revealing a dramatic improvement in source property constraints compared to the moderate foreground case with the same galaxy survey parameters. Across the test dataset, the 21cm galaxy cross-power spectrum (blue contours) achieves R^2 scores ranging from 0.81 to 0.86 across all four parameters, with $\text{PV} = 0.18 \pm 0.04$ relative to the prior. These constraints are broadly comparable to, though somewhat degraded relative to, those obtained with $M_{\text{h,min}} = 10^{10} M_{\odot}$ under moderate foregrounds ($\text{PV} = 0.11 \pm 0.02$, Section 4.4), demonstrating that recovering wedge modes can largely compensate for detecting fewer galaxies. Critically, the 21cm auto-power spectrum (grey contours) shows no improvement under optimistic foregrounds, with R^2 scores remaining between 0.09 and 0.51 and $\text{PV} = 0.73 \pm 0.03$. This asymmetric improvement reveals a fundamental physical distinction. The auto-power spectrum is inherently insensitive to source proper-

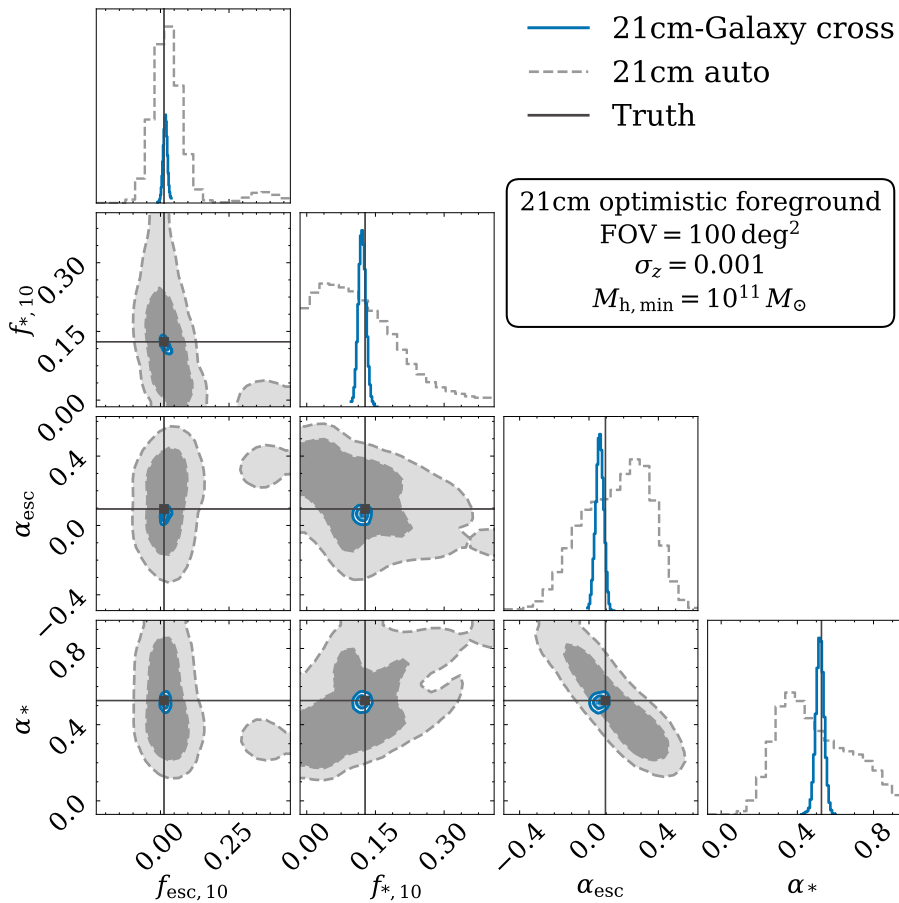


Fig. D.1. Inference of astrophysical source parameters from 21cm auto- and 21cm galaxy cross-power spectra for a randomly chosen sample assuming optimistic 21cm foreground removal. Corner plots show the marginalized posteriors for the four parameters controlling escape fraction and star formation efficiency: $\log_{10} f_{\text{esc},10}$, $\log_{10} f_{*,10}$, α_{esc} , and α_* . Blue, solid contours show cross-power constraints ($R^2 > 0.81$, posterior volumes $\sim 18\%$ across the entire test dataset). Grey, dashed contours show auto-power constraints, which are virtually uninformative ($R^2 < 0.51$, posterior volumes $\sim 73\%$ across the entire test dataset). Results assume the fiducial survey configuration with $\text{FOV} = 100 \text{ deg}^2$, spectroscopic redshift uncertainty $\sigma_z = 0.001$, and halo mass detection threshold $M_{\text{h,min}} = 10^{11} M_{\odot}$, but with an optimistic 21cm foreground model that recovers wedge modes down to the primary beam scale. The dramatic improvement in cross-power constraints compared to moderate foregrounds with this mass threshold (Section 4.4) demonstrates that recovering wedge modes can compensate for detecting fewer faint galaxies, offering a complementary observational pathway to constraining source properties. The persistent weakness of auto-power constraints reveals that additional 21cm modes do not break the fundamental degeneracies among source models that produce similar global ionization histories, highlighting the unique capability of cross-correlations to access source property information.

ties because different combinations of escape fractions and star formation efficiencies can produce similar global ionization histories and thus nearly identical power spectrum shapes. Recovering additional modes does not break these degeneracies. In contrast, the cross-power spectrum directly probes the correlation between ionized regions and galaxy positions, which depends sensitively on how efficiently individual galaxies produce ionizing photons. The large-scale modes recovered by optimistic foreground removal are precisely those that encode the strongest source-bubble correlations, explaining why cross-power constraints improve substantially while auto-power constraints remain weak.

This result establishes two complementary pathways to constraining reionization source properties through 21cm galaxy cross-correlations. Deep galaxy surveys detecting faint sources ($M_{\text{h,min}} \sim 10^{10} M_{\odot}$) can achieve tight constraints even under conservative 21cm foreground treatment. Alternatively, surveys targeting brighter galaxies ($M_{\text{h,min}} \sim 10^{11} M_{\odot}$), which are more readily accessible to planned facilities such as the WST and PFS, can reach similar precision if paired with aggressive 21cm foreground cleaning that successfully recovers wedge modes. This flexibility is crucial for practical observational programs, as it allows constraints on source properties to be achieved through improvements in either the galaxy survey sensitivity or the 21cm foreground removal capability. The dramatic contrast between cross-power and auto-power performance under optimistic foregrounds ($R^2 \sim 0.83$ versus $R^2 \sim 0.31$) underscores the unique role of cross-correlations in accessing source properties. While both observables benefit from increased sensitivity to the 21cm signal morphology, only the cross-power translates this into meaningful constraints on the parameters controlling where and

how efficiently ionizing photons escape from galaxies. This reinforces the conclusion of Section 4.4 that 21cm galaxy cross-correlations provide not merely complementary information to auto-power measurements, but essential information for parameters that remain degenerate in single-probe analyses.