Astronomy & Astrophysics

How informative are summaries of the cosmic 21 cm signal?

David Prelogović[*] and Andrei Mesinger

Scuola Normale Superiore, Piazza dei Cavalieri 7, 56125 Pisa, Italy

ABSTRACT

The cosmic 21 cm signal will bring data-driven advances to studies of Cosmic Dawn (CD) and the Epoch of Reionization (EoR). Radio telescopes such as the Square Kilometre Array (SKA) will eventually map the HI fluctuations over the first billion years – the majority of our observable Universe. With such large data volumes, it becomes increasingly important to develop "optimal" summary statistics, which will allow us to learn as much as possible about the CD and EoR. In this work we compare the astrophysical parameter constraining power of several 21 cm summary statistics, using the determinant of the Fisher information matrix, det F. Since we do not have an established "fiducial" model for the astrophysics of the first galaxies, we computed for each summary the distribution of det F across the prior volume. Using a large database of cosmic 21 cm light cones that include realizations of telescope noise, we compared the following summaries: (i) the spherically averaged power spectrum (1DPS), (ii) the cylindrically averaged power spectrum (2DPS), (iii) the 2D wavelet scattering transform (WST), (iv) a recurrent neural network (RNN) trained as a regressor; (v) an information-maximizing neural network (IMNN); and (vi) the combination of 2DPS and IMNN. Our best performing individual summary is the 2DPS, which provides relatively high Fisher information throughout the parameter space. Although capable of achieving the highest Fisher information for some parameter choices, the IMNN does not generalize well, resulting in a broad distribution across the prior volume. Our best results are achieved with the concatenation of the 2DPS and IMNN. The combination of only these two complimentary summaries reduces the recovered parameter variances on average by factors of ∼6.5–9.5, compared with using each summary independently. Finally, we point out that that the common assumption of a constant covariance matrix when doing Fisher forecasts using 21 cm summaries can significantly underestimate parameter constraints.

Key words. methods: data analysis – methods: statistical – cosmology: theory – dark ages, reionization, first stars

1. Introduction

The cosmic 21 cm signal, corresponding to the spin flip transition of the ground state of HI, provides a window on the first billion years of the Universe's evolution. This under-explored period witnessed fundamental cosmic milestones, including the Cosmic Dawn (CD) of the first galaxies and the final phase change of our Universe: the Epoch of Reionization (EoR). Current radio interferometers are setting increasingly tight upper limits on its power spectrum (PS; e.g., Trott et al. 2020; Mertens et al. 2020; Gehlot et al. 2020; HERA Collaboration 2023; Munshi et al. 2024), while the upcoming Square Kilometre Array (SKA)[1] promises to provide a 3D image of the EoR in the next decade(s) (e.g., Koopmans et al. 2015; Mesinger 2020). The unknown properties of the first galaxies and intergalactic medium (IGM) structures are encoded in the timing and morphology of the cosmic 21 cm signal (e.g., Pritchard & Furlanetto 2007; McQuinn & O'Leary 2012; Visbal et al. 2012; Pacucci et al. 2014).

The only statistically robust way to infer these properties is through Bayesian inference: sampling from theory and comparing forward models to observations. But what should we use to compare the forward model to the data? In other words, how should we construct the likelidhood? The cosmic 21 cm signal is intrinsically a 3D light cone map. Performing inference directly on the light cone would be incredibly challenging due to the high dimensionality of the data. Effectively, at each frequency the SKA should obtain a sky map comparable to current cosmic microwave back-

ground maps, but with the full light cone including thousands of such frequency slices (e.g., Loeb & Zaldarriaga 2004)[2]. Instead, the data are compressed into a summary statistic. Summary statistics usually involve some form of averaging, which increases the signal-to-noise ratio (S/N) and can motivate assuming an approximately Gaussian likelihood when performing inference (e.g., Greig & Mesinger 2015; Gazagnes et al. 2021; Watkinson et al. 2022)[3].

The most common choice of a 21 cm summary statistic is the spherically averaged power spectrum (1DPS). The 1DPS maximizes the S/N of an interferometric observation – a primary concern in the run-up to a first detection (e.g., Greig & Mesinger 2018; Trott et al. 2020; Mertens et al. 2020;

---

[*] Corresponding author; david.prelogovic@sns.it
[1] https://www.skatelescope.org

[2] Inference on high-dimensional light cones is still possible in some cases, for example galaxy large-scale structure maps (e.g., Kitaura & Enßlin 2008; Jasche & Kitaura 2010; Jasche & Wandelt 2012, 2013; Leclercq et al. 2017; McAlpine et al. 2022; Dai & Seljak 2022; Bayer et al. 2023). Although such studies still compress the full galaxy observations, only treating them as biased tracers of the matter field, they are able to recover the phase information of our Universe's initial conditions (so-called constrained realizations). It remains to be seen if such studies can be extended to the EoR and CD (though see Zhao et al. 2023b for a recent proof-of-concept study).

[3] The Gaussianity of summaries constructed by averaging can be loosely motivated by appealing to the Central Limit Theorem. However, the theorem holds only for independent, identically distributed random variables. Unfortunately, we only have a single observable Universe; thus summaries of cosmological datasets (e.g., the PS) often involve binning over correlated (i.e., not independent) and/or differently evolving fields (i.e., not identically distributed), resulting in non-Gaussian likelihoods (e.g., Prelogović & Mesinger 2023).

HERA Collaboration 2023). Another common summary choice is the cylindrically averaged power spectrum (2DPS), which includes additional information on the anisotropy of the cosmic 21 cm signal (e.g., Bharadwaj & Ali 2005; Barkana & Loeb 2006; Datta et al. 2012; Mao et al. 2012). Fourier space is also a natural basis for isolating the cosmic signal from instrumental effects and foregrounds, which primarily reside in a wedge region in the 2DPS (e.g., Morales et al. 2012; Vedantham et al. 2012; Trott et al. 2012; Parsons et al. 2014; Liu et al. 2014a,b; Murray & Trott 2018). However, the PS ignores Fourier phases, which can encode significant information for a highly non-Gaussian signal such as the 21 cm signal from the EoR and CD (e.g., Bharadwaj & Pandey 2005; Mellema et al. 2006; Shimabukuro et al. 2016; Majumdar et al. 2018; Watkinson et al. 2019). Because of this, several studies have also explored non-Gaussian statistics, such as the bispectrum (Shimabukuro et al. 2016; Majumdar et al. 2018; Watkinson et al. 2019, 2022), the trispectrum (Cooray et al. 2008; Flöss et al. 2022), morphological spectra (Gazagnes et al. 2021), the bubble size distribution (Lin et al. 2016; Giri et al. 2018; Shimabukuro et al. 2022; Doussot & Semelin 2022; Lu et al. 2024), one-point statistics (Harker et al. 2009; Watkinson & Pritchard 2015; Kittiwisit et al. 2022), genus topology (Hong et al. 2014), wavelet scattering transform (WST; Greig et al. 2022, 2023; Zhao et al. 2023a), Minkowski functionals and tensors (Gleser et al. 2006; Yoshiura et al. 2017; Chen et al. 2019; Kapahtia et al. 2019; Spina et al. 2021), and Betti numbers (Giri & Mellema 2021; Kapahtia et al. 2021).

Is there an "optimal" choice of summary statistic? One way to define optimal is by how tightly we can recover cosmological and astrophysical parameters. This can be quantified through the Fisher information matrix (Fisher 1935). The Fisher information is commonly used to make forecasts using predetermined summary statistics, or even to define the summary statistic itself. Optimal summaries (algorithms) that are based on the Fisher matrix include Karhunen-Loéve methods and massively optimized parameter estimation and data compression for linear transformations (e.g., Tegmark et al. 1997; Heavens et al. 2000), as well as nonlinear generalizations like information-maximizing neural networks (IMNNs; Charnock et al. 2018) and fishnets (Makinen et al. 2023). However, these summary statistics come at the cost of interpretability, making the compression more of a "black box" compared to physically motivated and easy-to-interpret summaries like the PS[4].

It is also important to note that, unlike Λ cold dark matter (ΛCDM), there is no obvious or unique choice for parameterizing the astrophysics of early galaxies whose radiation determines the cosmic 21 cm signal. Nor do we have a good idea of a "fiducial" model. Thus, there is no guarantee that a summary statistic that maximizes Fisher information at a single parameter value for a single model parametrization would generalize well to other parameters and models.

Here we systematically compare the information content of several common 21 cm summaries on the basis of their Fisher information. These include: (i) the 1DPS, (ii) the 2DPS, (iii) the 2D WST, (iv) a recurrent neural network (RNN) trained

as a regressor; and (v) an IMNN[5]. Several previous studies compared the constraining power of 21 cm summaries (e.g., Gazagnes et al. 2021; Watkinson et al. 2022; Greig et al. 2022; Zhao et al. 2023b; Hothi et al. 2024). Here we explore a broader range of summary statistics, introducing IMNNs to the field. Unlike previous 21 cm Fisher studies, we do not make the simplifying assumption of a constant covariance matrix. Most importantly, we compare the summary statistics across a broad parameter space, instead of choosing only one or two fiducial values to make a mock 21 cm observation. This is very important for verifying the robustness of the summary since we do not have a fiducial astrophysical model for the first galaxies.

The paper is organized as follows. In Sect. 2 we define the information content of a summary through the Fisher matrix and introduce the Fisher information distribution. In Sect. 3 we introduce our 21 cm simulator and explain the summaries we consider in detail. In Sect. 4 we discuss our database of prior samples for evaluating the Fisher information. In Sect. 5 we present our results, quantifying the most informative summaries as well as the error introduced by the common assumption of a constant covariance matrix. Finally, we present concluding remarks and future prospects in Sect. 6. All quantities are quoted in comoving units, and we assume a standard ΛCDM cosmology: $(\Omega_\Lambda, \Omega_M, \Omega_b, n, \sigma_8, H_0)$ = $(0.69, 0.31, 0.048, 0.97, 0.81, 68 \text{ km s}^{-1}\text{Mpc}^{-1})$, consistent with the results from Planck Collaboration VI (2020).

2. The Fisher matrix as a measure of information content

How do we measure the information content of a summary? If we have a theoretical model with a given parametrization, we can see how sensitive the summary is to changes in the astrophysical/cosmological parameters around some fiducial value. Specifically, we can calculate the Fisher matrix (e.g., Spall 2005):

$$
\begin{aligned}
\boldsymbol{F}(\boldsymbol{\theta}^*)_{mn} &= \int d\boldsymbol{d}\, P(\boldsymbol{d}|\boldsymbol{\theta}^*)\ \frac{\partial}{\partial\boldsymbol{\theta}_m}\ln P(\boldsymbol{d}|\boldsymbol{\theta}^*)\cdot\frac{\partial}{\partial\boldsymbol{\theta}_n}\ln P(\boldsymbol{d}|\boldsymbol{\theta}^*) \\
&= \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}_m}\ln P(\boldsymbol{d}|\boldsymbol{\theta})\cdot\frac{\partial}{\partial\boldsymbol{\theta}_n}\ln P(\boldsymbol{d}|\boldsymbol{\theta})\middle|\boldsymbol{\theta}^*\right],
\end{aligned}
\tag{1}
$$

where $\boldsymbol{d}$ is the data summary (cf. Fig. 1), $P(\boldsymbol{d}|\boldsymbol{\theta})$ is the likelihood, and $\mathbb{E}$ denotes the expectation value over the likelihood (i.e., an empirical average over many realizations) evaluated at a given point in parameter space $\boldsymbol{\theta}^*$. The Fisher information matrix provides the maximum constraining power, known as the Cramér-Rao bound (Rao 1992; Cramér 1999): $\text{Var}(\hat{\boldsymbol{\theta}}_m) \geq (\boldsymbol{F}^{-1})_{mm}$, where $\hat{\boldsymbol{\theta}}$ is an unbiased estimator of the parameters. In the multivariate case, one can equivalently write:

$$
\text{Cov}(\hat{\boldsymbol{\theta}}) \geq \boldsymbol{F}^{-1},
\tag{2}
$$

where the matrix inequality is interpreted as $\text{Cov}(\hat{\boldsymbol{\theta}}) - \boldsymbol{F}^{-1}$ being a positive semi-definite matrix. One can then prove the following inequality (see Appendix A):

$$
\det\text{Cov}(\hat{\boldsymbol{\theta}}) \geq \det \boldsymbol{F}^{-1}.
\tag{3}
$$

Therefore, the determinant of the Fisher matrix measures the tightest parameter volume that can be constrained by the data.

---

[4] An interesting question is why would we even expect a truly optimal (lossless) compression to exist? The answer to this can be motivated through the manifold hypothesis (Fefferman et al. 2016), which states that physical probability distributions of the data often lie on a low-dimensional manifold. Optimal (lossless) compression would then "extract" such manifolds from the original high-dimensional data. The above-mentioned algorithms do not attempt this directly, but aim at making compression as optimal as possible, given certain assumptions.

[5] Note that the dimensionality of the PS (1D or 2D) labels the dimensionality of the summary, while for the WST, 2D labels the dimensionality of the data on which it is computed. In our case, those are 2D sky-plane images.

From now on, we refer to det $\boldsymbol{F}(\boldsymbol{\theta}^*)$ as simply the Fisher information around $\boldsymbol{\theta}^*$ – the higher its value, the more constraining the summary is around $\boldsymbol{\theta}^*$.

2.1. Gaussian approximation

In almost all cosmological scenarios, we do not explicitly know the full likelihood function $P(\boldsymbol{d}|\boldsymbol{\theta})$, nor its "score", $\nabla_{\boldsymbol{\theta}} \ln P(\boldsymbol{d}|\boldsymbol{\theta})$. For performing quick forecasts, it is instead common to approximate the likelihood as a Gaussian in data space $P(\boldsymbol{d}|\boldsymbol{\theta}) \approx \mathcal{N}(\boldsymbol{d}|\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, where the mean $\boldsymbol{\mu}(\boldsymbol{\theta})$ and covariance $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ are general functions of the parameter space.

Assuming a Gaussian likelihood, one can compute the integral in Eq. (1) analytically (Appendix B; see also Tegmark et al. 1997; Vogeley & Szalay 1996):

$$
\begin{aligned}
\boldsymbol{F}_{mn} &\equiv \boldsymbol{F}_{\mu,mn} + \boldsymbol{F}_{\Sigma,mn} \\
&= \frac{\partial \boldsymbol{\mu}^T}{\partial \theta_m} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \theta_n} + \frac{1}{2} \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_m} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_n} \right].
\end{aligned} \tag{4}
$$

Here the first (second) term measures how the mean (covariance) of the data summary changes with respect to the parameters. Cosmological Fisher forecasts usually ignore one of the two terms (most often the second term $\boldsymbol{F}_{\Sigma}$), effectively fixing either the mean or the covariance matrix to some fiducial value. However, we will see below that both terms are non-negligible for some summaries (e.g., Carron 2013).

2.2. Finite differencing

To numerically evaluate Eq. (4), we grouped the terms into "fixed" ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{-1}$) and "differentiated" ($\nabla_{\boldsymbol{\theta}} \boldsymbol{\mu}$, $\nabla_{\boldsymbol{\theta}} \boldsymbol{\Sigma}$). We calculated fixed terms from realizations of the summary around the same parameter value $\boldsymbol{\theta}^*$. To compute the mean, we can use the maximum likelihood estimator (MLE):

$$
\boldsymbol{\mu}(\boldsymbol{\theta}^*) = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{d}(\boldsymbol{\theta}^*)_i, \tag{5}
$$

where the index i labels different realizations of initial conditions and other sources of stochasticity (e.g., telescope noise). It is known that such an estimator of the mean is unbiased. However, for the covariance matrix the situation is more complex, as the MLE

$$
\boldsymbol{\Sigma}_{\mathrm{MLE}}(\boldsymbol{\theta}^*) = \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{d}(\boldsymbol{\theta}^*)_i - \boldsymbol{\mu}(\boldsymbol{\theta}^*))(\boldsymbol{d}(\boldsymbol{\theta}^*)_i - \boldsymbol{\mu}(\boldsymbol{\theta}^*))^T \tag{6}
$$

is biased. One can show that the estimators of the covariance and its inverse can be unbiased in the following way (e.g., Hartlap et al. 2007):

$$
\boldsymbol{\Sigma}(\boldsymbol{\theta}^*) = \frac{N}{N-1} \boldsymbol{\Sigma}_{\mathrm{MLE}}(\boldsymbol{\theta}^*), \tag{7}
$$

$$
\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}^*) = \frac{N-D-2}{N-1} \boldsymbol{\Sigma}_{\mathrm{MLE}}^{-1}(\boldsymbol{\theta}^*), \tag{8}
$$

where D is dimensionality of the summary $\boldsymbol{d}$. The correction prefactor for the inverse of the covariance is particularly important when D is close to the number of samples N. Without it, both $\boldsymbol{F}_{\mu}$ and $\boldsymbol{F}_{\Sigma}$ would end up over-confident, with the effect on $\boldsymbol{F}_{\Sigma}$ being quadratic (see Eq. (4)).

A similar approach is used to estimate the "differentiated parameters". In the case in which the simulator is not differentiable, we can use finite differencing to estimate gradients of the mean and covariance:

$$
\frac{\delta \boldsymbol{d}(\boldsymbol{\theta}^*)_i}{\delta \boldsymbol{\theta}_m} \equiv \frac{\boldsymbol{d}(\boldsymbol{\theta}^* + \delta\theta_m/2)_i - \boldsymbol{d}(\boldsymbol{\theta}^* - \delta\theta_m/2)_i}{\delta\theta_m}, \tag{9}
$$

$$
\left.\frac{\partial \boldsymbol{\mu}}{\partial \theta_m}\right|_{\boldsymbol{\theta}^*} = \frac{1}{M} \sum_{i=1}^{M} \frac{\delta \boldsymbol{d}(\boldsymbol{\theta}^*)_i}{\delta\theta_m}, \tag{10}
$$

$$
\left.\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_m}\right|_{\boldsymbol{\theta}^*} = \frac{2}{L-1} \sum_{i=1}^{L} \left(\frac{\delta \boldsymbol{d}(\boldsymbol{\theta}^*)_i}{\delta\theta_m} - \left.\frac{\partial \boldsymbol{\mu}}{\partial \theta_m}\right|_{\boldsymbol{\theta}^*} \right) (\boldsymbol{d}(\boldsymbol{\theta}^*)_i - \boldsymbol{\mu}(\boldsymbol{\theta}^*))^T. \tag{11}
$$

Here, $\delta\theta_m$ represents a small change in the parameter over which the derivative is computed and $\delta\boldsymbol{d}(\boldsymbol{\theta}^*)_i/\delta\theta_m$ corresponds to the numerical derivative of a given summary realization (i.e., fixing all sources of stochasticity).

2.3. The distribution of the Fisher information over parameter space

Below we show examples of common 21 cm summaries evaluated at a fiducial set of parameters, $\boldsymbol{\theta}_{\mathrm{fid}}$. Unfortunately, the cosmic 21 cm signal does not yet have a well-defined fiducial model and our choice of $\boldsymbol{\theta}_{\mathrm{fid}}$ is only a "best guess". Therefore, to properly assess the quality of a particular summary, we sampled many points $\boldsymbol{\theta}^*$ from the prior $P(\boldsymbol{\theta})$, calculating the Fisher information det $\boldsymbol{F}(\boldsymbol{\theta}^*)$ at each sample. We used this prior-weighted distribution of Fisher information as the main metric for assessing the quality of a given summary[6].

3. The cosmic 21 cm signal and its summary statistics

For a given sample of astrophysical parameters, $\boldsymbol{\theta}^*$, we computed a realization of mock data using the following steps (illustrated in Fig. 1):

1. Sampling cosmological initial conditions, we simulated a realization of the 21 cm lightcone using the `21cmFAST` code.
2. We subtracted the mean from each frequency slice, mimicking an interferometric observation.
3. We simulated a 1000 h tracked scan with the SKA1-low telescope, including the corresponding uv coverage and realization of the instrument noise.
4. We compressed the resulting light cone into a given summary observation.

We elaborate on these steps below.

The cosmic 21 cm signal was simulated using the semi-numerical code, `21cmFASTv3`[7] (Mesinger & Furlanetto 2007; Mesinger et al. 2011; Murray et al. 2020). In brief, we used the Park et al. (2019) galaxy model, varying five parameters that characterize the unknown UV and X-ray properties of high-z galaxies through scaling relations:

– $f_{*,10}$ – the fraction of galactic gas in stars, normalized at the halo mass of $10^{10} M_{\odot}$. The stellar to halo mass relation of the faint galaxies that drive reionization is well described by a power law: $M_*/M_h = f_{*,10}(M_h/10^{10}M_{\odot})^{0.5}(\Omega_b/\Omega_m)$ (see Park et al. 2019 and references therein).

---

[6] Note that the prior-weighted distribution of the Fisher information is closely related to the Mutual Information $I(\boldsymbol{x}, \boldsymbol{\theta}) = D_{\mathrm{KL}}(p(\boldsymbol{x}, \boldsymbol{\theta}) \| p(\boldsymbol{x}) p(\boldsymbol{\theta}))$, where D_{KL} is Kullback-Liebler divergence (Brunel & Nadal 1998). Mutual Information can also be used to asses the quality of a summary (e.g., Sui et al. 2023).
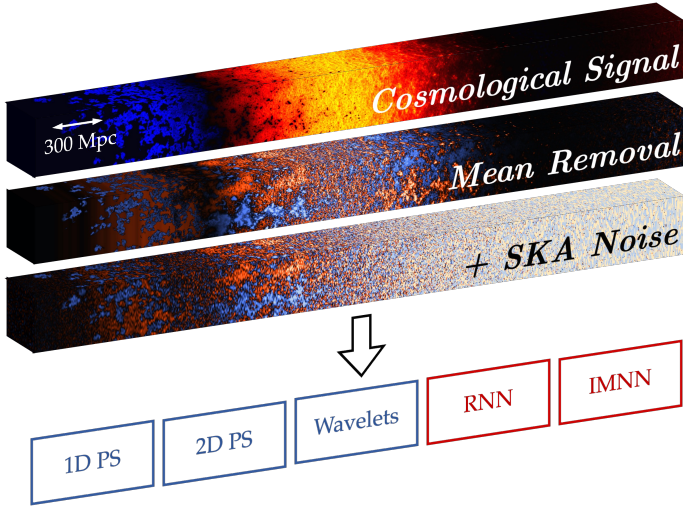
[7] https://github.com/21cmfast/21cmFAST

Fig. 1. Schematic illustrating our pipeline for producing data summaries for a given astrophysical parameter combination (adapted from Prelogović & Mesinger 2023). Starting from a realization of the cosmological signal simulated with `21cmFAST`, we remove the mean at each redshift and add a telescope noise realization corresponding to a 1000 h observation with SKA1-Low. The resulting light cone is compressed into different summaries. Explicitly defined summaries are marked in blue, while neural network summaries are marked in red. This realization is computed at the fiducial parameter set, which was used only for IMNN training and visualizations of the summaries. See the main text for more details.

- $f_{\rm esc,10}$ – the ionizing UV escape fraction, normalized at the halo mass $10^{10} M_\odot$. Again, the (mean) escape fraction is assumed to be a power law with halo mass, and here we fixed its index to $\alpha_{\rm esc} = -0.5$ (e.g., Qin et al., in prep.).
- $M_{\rm turn}$ – the characteristic host halo mass below which galaxies are inefficient at forming stars, due to slow gas accretion, supernovae feedback, and/or photo-heating.
- E_0 – the characteristic X-ray energy below which photons are absorbed by the interstellar medium (ISM) of the host galaxy.
- $L_{X<2\rm keV}$/SFR – the X-ray soft band (in the energy range $E_0 - 2$ keV) luminosity per star formation rate (SFR), in units of erg s^{-1} keV^{-1} M$_\odot^{-1}$ yr. The SFR is taken to be: $\dot{M}_* = M_*/(0.5\,H(z)^{-1})$, where the Hubble time, $H(z)^{-1}$, also scales with the dynamical time at the virial radius.

Our fiducial, $\boldsymbol{\theta}_{\rm fid}$ parameter vector corresponds to the following: $\log_{10} f_{*,10} = -1.3$, $\log_{10} f_{\rm esc} = -1$, $\log_{10} M_{\rm turn}[M_\odot] = 8.7$, $\log_{10} L_{X<2\rm keV}$/SFR[erg s^{-1} keV^{-1} M$_\odot^{-1}$ yr] $= 40$, $E_0[\rm keV] = 0.5$. This particular choice is only used when training the IMNN, and for the visualizations of the summaries below. More details on the model and its motivations are provided in Park et al. (2019) and references therein.

To simulate the observation, we followed the procedure described in Prelogović et al. (2022), Prelogović & Mesinger (2023; cf. Fig. 1). In brief, we calculated the 21 cm signal from redshifts 30 to 5, on a 300 Mpc coeval cube, with 1.5 Mpc resolution. The snapshots of the evolved coeval cube were then stacked to produce the final light cone of the signal. After the subtraction of the mean for every sky-plane slice, we added thermal noise corresponding to a a 6 h/day, 10 s integration time tracked scan with SKA1-Low antenna configuration for a 1000 h observation in total. For this task we used `tools21cm`[8], where a separate

---

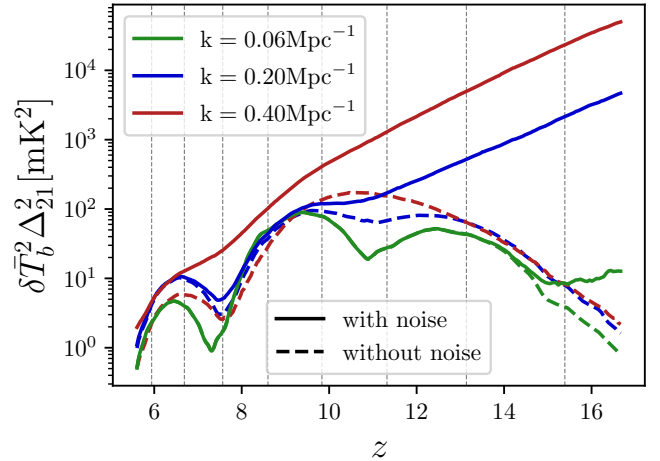[8] https://github.com/sambit-giri/tools21cm

Fig. 2. 1DPS for the fiducial light cone shown in Figure 1. Different colors correspond to different wave modes, while solid (dashed) lines correspond to light cones with (without) noise. Comparing the solid and dashed lines, we see that the signal is noise dominated at higher k modes and redshifts (see also Fig. 1). All power spectra are computed from comoving cubes extracted from the light cone, centered on the redshifts indicated by the vertical dashed lines.

UV coverage and noise is calculated for each sky-plane slice. Before computing each summary, we additionally smoothed the light cone with a box-car filter obtaining a final resolution of 6 Mpc. This final step does not impact our results as it is comparable to the SKA1-Low beam, but is needed in order to minimize GPU memory usage for certain summaries (see below for more details). We also note that in this work we made the optimistic assumption of perfect foreground removal, thus exploring the maximum future potential of our set of summary statistics.

The resulting 3D light cone was compressed into a given summary statistic. We explore the following summaries in this work:
1DPS – spherically averaged 1D power spectrum,
2DPS – cylindrically averaged 2D power spectrum,
Wavelets – 2D wavelet scattering transform,
RNN – recurrent neural network,
IMNN – information-maximizing neural network.
2DPS + IMNN – concatenation of the 2DPS and IMNN summaries.
The first four are explicitly defined while the following two are "learned" by neural networks (NNs). The last, 2DPS + IMNN, is a combination of the two individual summaries. We describe them in detail below.

3.1. Spherically averaged (1D) power spectrum

As discussed in the introduction, the 1DPS, is the most common choice of a summary statistic. Despite the cosmic 21 cm signal being non-Gaussian, the 1DPS is a natural choice when seeking to maximize the S/N of an interferometric first detection.

The 1DPS of the 21 cm signal $\delta T_b(\boldsymbol{x}, z)$, where $\boldsymbol{x}$ is the sky-plane coordinate and z the redshift, is defined as

$$\delta \bar{T}_b^2 \Delta_{21}^2(k, z') \equiv \frac{k^3}{2\pi^2 V} \left\langle \left| \delta T_b(\boldsymbol{k}, z') - \delta \bar{T}_b(z') \right|^2 \right\rangle_k . \quad (12)$$

The $\delta T_b(\boldsymbol{k}, z')$ is calculated from the Fourier transforms of mean-subtracted light cone segments centered around z'. Here we calculated the 1DPS in three log-spaced k-bins, computed at eight different z'. In Fig. 2 we show the 1DPS of the fiducial light cone

(see Fig. 1). Solid curves correspond to the full signal: cosmic + noise, while the dashed curves are only the cosmic signal. We see the usual three peaked redshift evolution of the large-scale power at $k \sim 1 \, \mathrm{Mpc}^{-1}$, tracing fluctuations in the IGM ionized fraction, the IGM temperature and Lyman alpha background (e.g., Pritchard & Furlanetto 2007; Mesinger et al. 2014). By comparing the solid and dashed curves, we see that the signal is noise dominated at small scales and early times. Indeed, $k \geq 0.4 \, \mathrm{Mpc}^{-1}$ remains noise dominated virtually at all redshifts ($z > 7$), for this fiducial model and choice of 1000 h integration time.

3.2. Cylindrical (2D) power spectrum

The cylindrically averaged 2DPS distinguishes between sky-plane ($k_\perp$) and line-of-sight ($k_\parallel$) modes. Redshift space distortions (e.g., Bharadwaj & Ali 2005; Barkana & Loeb 2005) and light cone evolution (e.g., Greig & Mesinger 2018; Mondal et al. 2018) result in an anisotropic cosmic signal. Therefore, the 2DPS should encode more physical information compared to the 1DPS. Moreover, the instrument/foregrounds are better characterized in the 2DPS (e.g., see the review in Liu et al. 2014b), making it a natural summary observable for preliminary, low S/N measurements.

The 2DPS is defined as

$$\delta \bar{T}_b^2 \Delta_{21}^2(k_\perp, k_\parallel, z') \equiv \frac{k_\perp^2 \, k_\parallel}{4\pi^2 V} \left\langle \left| \delta T_b(\mathbf{k}, z') - \delta \bar{T}_b(z') \right|^2 \right\rangle_{k_\perp, k_\parallel}, \quad (13)$$

where the Fourier transform is performed on the same light cone chunks as in the 1D case, with the difference being that the expectation value is calculated over ($k_\perp, k_\parallel$) bins. Here we picked 3×3 log-spaced bins, aligned to the 1D case as close as possible. The 2DPS for the fiducial light cone (see Fig. 1) is shown in Fig. 3. The panels correspond to the mean redshifts of the light cone chunks, coinciding with the ones of the 1DPS (see Fig. 2). The columns correspond to the cases with and without noise, as labeled at the top. We can verify that both the cosmic signal and the noise are anisotropic. For the cosmic signal, this is most evident for $z \lesssim 10$ where the noise is subdominant. For the noise, the same can be clearly seen for higher redshifts, where the noise dominates the signal at high $k_\parallel$ much more than in high $k_\perp$.

3.3. Wavelet scattering transform

The wavelet transform (Gabor 1946; Goupillaud et al. 1984; Trott et al. 2012) and WST (Mallat 2011) are based on convolutional filters designed on a Fourier basis with an additional Gaussian envelope along the frequency direction. As such, they capture local features in both image and Fourier domains, which is often relevant in audio and image processing. Higher-order convolutions combined with NNs (wavelet-based convolutional NNs) have been very successful for such purposes (Bruna & Mallat 2012; Sifre & Mallat 2013; Anden & Mallat 2014). The WST has been extensively explored in cosmological (e.g., Cheng et al. 2020) and 21 cm (e.g., Greig et al. 2022, 2023; Zhao et al. 2023a; Hothi et al. 2024) analyses.

Here we briefly introduce how the WST is defined. It represents a convolution of an input image $I(\mathbf{x})$ by a set of wavelet (here specifically Morlet) filters. They are characterized by the physical scale j and rotation moment l. The actual physical scale of the filter $\psi^{j,l}$ will then be 2^j pixels, and its orientation angle $l \cdot \pi / L$. The wavelet coefficients correspond to the average of the convolved images, to which we refer as simply wavelets. We note that in this work we restricted ourselves to the 2D WST

(see, e.g., Zhao et al. 2023a for recent work using the 3D transform).

The first- and second-order wavelets can be written as

$$s_1^{j_1, l_1} = \langle |I(\mathbf{x}) * \psi^{j_1, l_1}| \rangle, \quad (14)$$

$$s_2^{j_1, l_1, j_2, l_2} = \langle \||I(\mathbf{x}) * \psi^{j_1, l_1}| * \psi^{j_2, l_2}| \rangle. \quad (15)$$

One can see the similarity between first-order wavelets and the PS by writing $\mathrm{PS}(\mathbf{k}) = \langle |I(\mathbf{x}) * \psi'|^2 \rangle$, with $\psi' = e^{-i\mathbf{k} \cdot \mathbf{x}}$. To extract only cylindrically averaged information imprinted in the 21 cm signal (analogous to calculating 2DPS in a Fourier basis), one can average coefficients over the rotations:

$$S_1^{j_1} = \langle s_1^{j_1, l_1} \rangle_{l_1}, \quad (16)$$

$$S_2^{j_2} = \langle s_2^{j_1, l_1, j_2, l_2} \rangle_{l_1, l_2}. \quad (17)$$

It is important to mention that this procedure removes anisotropic information. Alternative definitions such as the reduced WST (Allys et al. 2019) are able to reduce the coefficient dimensionality, while keeping anisotropy. For further information about the WST and wavelet coefficients we refer the reader to Cheng et al. (2020) and Greig et al. (2022, 2023).

Our setup for calculating wavelets is the following. We first selected a series of eight images, centered at the same redshifts as for the power spectra. For each image we calculated the first- and second-order wavelets and combined them for the final summary. As the WST scale ($2^J \times 2^J$) has to be smaller than the sky-plane images (50×50), we set $J = 4$. Furthermore, we fixed $L = 4$ (Cheng et al. 2020; Greig et al. 2022). To compute convolutions, we used the `scattering_transform`[9] package (Cheng et al. 2020). For each redshift slice, we averaged the first- and second-order coefficients computed on a rolling 32×32 window. Our final WST summary thus consists of $J = 4$ S_1 and $J(J+1)/2 = 10$ S_2 coefficients for each of the eight redshift slices, for a total of 112 numbers.

In Figure 4, we show the redshift evolution of first-order wavelet coefficients (S_1; see Eqs. (14), (16)). Solid (dashed) lines correspond to coefficients with (without) telescope noise. We see the same qualitative trends as in Fig. 2, with lower J modes (i.e., smaller scales) more noise dominated. As a result, these low J modes contribute less to the total Fisher information (Hothi et al. 2024).

3.4. Recurrent neural network

A RNN is a type of NN architecture designed to efficiently encode sequential or time-evolving information. In our previous work (Prelogović et al. 2022), we used RNNs along the redshift axis in combination with convolutional neural networks (CNNs) along the sky-plane, to capture the anisotropic evolution of the cosmic 21 cm signal along the light cone. Specifically, we used a long short-term memory (Hochreiter & Schmidhuber 1997; Shi et al. 2015) RNN, which is more stable to train compared with older RNN versions (see, e.g., the review in Schmidt 2019). Using RNNs resulted in a more accurate regression of astrophysical parameters, compared to only using CNNs (e.g., Gillet et al. 2019; La Plante & Ntampaka 2019; Mangena et al. 2020; Kwon et al. 2020; Zhao et al. 2022a; Heneka 2023).

In this work, we used the flagship model from Prelogović et al. (2022), *SummaryRNN*, together with the exact weights from that work. The model has been trained in a supervised manner as a regressor. The database used for

---

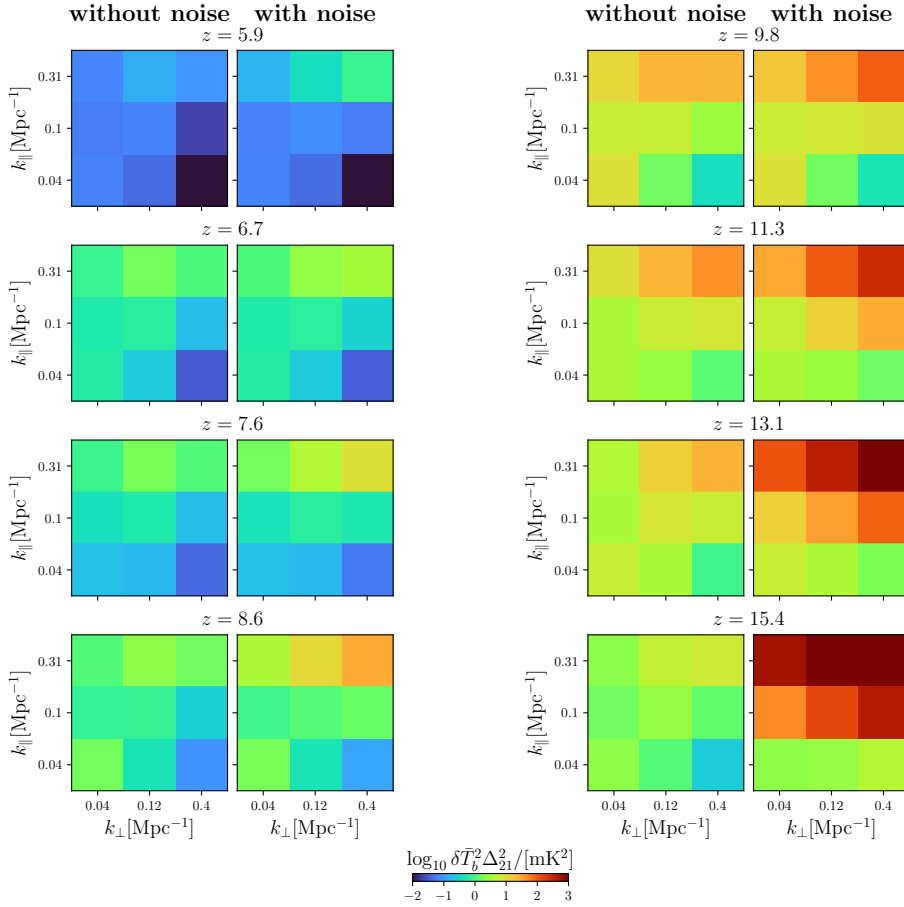[9] https://github.com/SihaoCheng/scattering_transform

Fig. 3. 2DPS of the fiducial model shown in Fig. 1. Different columns show the 2DPS for the pure cosmic signal and including thermal noise, as labeled on the top. Each pair corresponds to a different central redshift, coinciding with those used in the 1DPS summary (see Fig. 2). Furthermore, the bins between 1DPS and 2DPS coincide as much as possible. As in the 1DPS case, high-k modes and high redshifts are noise dominated.
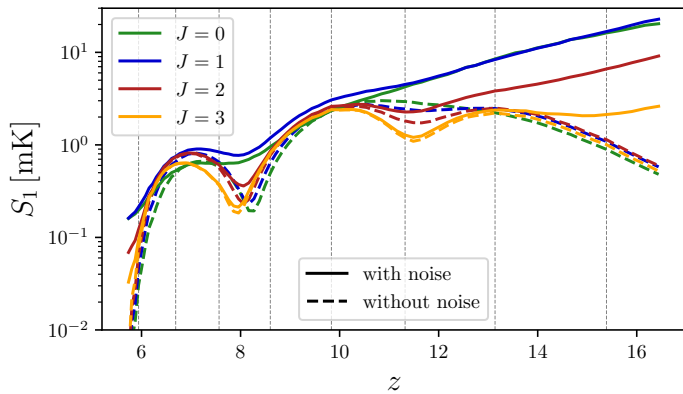

Fig. 4. First-order wavelet coefficients (Eqs. (14) and (16)) for the fiducial light cone shown in Figure 1. Different colors correspond to different J modes, while solid (dashed) lines correspond to the case with (without) noise. Comparing the solid and dashed lines, we see that the signal is noise dominated at smaller scales (smaller J) and higher redshifts (see also Fig. 1).

training had the same telescope simulator, but used an older version of the `21cmFAST` simulator with a different astrophysical parametrization.

Our RNN summary in this work is the 32-dim output from the first dense layer of the network (see Table B3 in Prelogović et al. 2022, where convolutional and recurrent layers are followed by dense layers, ending with prediction of four parameters). Using such a higher-dimensional layer instead of the final parameter regressor output (e.g., Zhao et al. 2022a),

makes the summary more sensitive to the general light cone features instead of the specific model and/or parametrization used to create the training set. This is similar to so-called transfer learning (e.g., Jiang et al. 2022), where one uses a NN that has been pretrained on a different domain (i.e., database) and retrains its last layers on a new domain. As our original RNN takes as input 25×25 sky-plane slices, we rolled a 25×25 window over each redshift slice, averaging the corresponding dense layer outputs to obtain the final RNN summary.

3.5. Information-maximizing neural network

An IMNN (Charnock et al. 2018)[10] is an unsupervised learning algorithm, where the NN is specifically trained to maximize the Fisher information $\det \boldsymbol{F}$ of the summary compression, at a single fixed point in parameter space $\boldsymbol{\theta}_{\mathrm{fid}}$. IMNNs have been previously applied to galaxy large-scale structure maps (e.g., Makinen et al. 2022) and catalog-free modeling of galaxy types (Livet et al. 2021). Here we applied them for the first time to 21 cm maps of the EoR and CD. In what follows, we outline the main ingredients of the algorithm and the training procedure.

Taking some NN architecture (in our case a simple CNN), one compresses the light cone, $\boldsymbol{l}$, into a summary vector, $\boldsymbol{d}$,

$$\boldsymbol{d} = \mathrm{NN}(\boldsymbol{l}) \,. \tag{18}$$

The dimensionality of the summary is equal to the parameter space dimensionality (D). This is motivated by the fact that, for a given model parametrization, the "score" of the real likelihood

---

[10] https://github.com/tomcharnock/IMNN

function, $\nabla_\theta \ln P(l|\theta)$, is sufficient for optimal parameter recovery, and thus the dimensionality of the summary that maximizes information around one fiducial point can be the same as the dimensionality of the parameter space (e.g., Heavens et al. 2000; Charnock et al. 2018).

We trained the IMNN using the following loss function (cf. Charnock et al. 2018; Makinen et al. 2021):

$$F_{\text{IMNN}}(\theta_{\text{fid}})_{mn} \equiv \left. \frac{\partial \boldsymbol{\mu}^T}{\partial \boldsymbol{\theta}_m} \right|_{\theta_{\text{fid}}} \left. \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}_n} \right|_{\theta_{\text{fid}}} \tag{19}$$

$$L_{\mathcal{F}}(\boldsymbol{\Sigma}(\theta_{\text{fid}})) \equiv \|\boldsymbol{\Sigma}(\theta_{\text{fid}}) - \mathbb{1}\|_{\mathcal{F}} \tag{20}$$

$$L_2(W_{\text{NN}}) \equiv \|W_{\text{NN}}\|^2 \tag{21}$$

$$\begin{aligned} \mathcal{L} = &- \ln \det \boldsymbol{F}_{\text{IMNN}} \\ &+ \lambda \, L_{\mathcal{F}}(\boldsymbol{\Sigma}) \tanh L_{\mathcal{F}}(\boldsymbol{\Sigma}) \\ &+ \lambda_W \, L_2(W_{\text{NN}}), \end{aligned} \tag{22}$$

where

$$\boldsymbol{\mu}(\theta_{\text{fid}}) = \frac{1}{N} \sum_{i=1}^{M} \boldsymbol{d}(\theta_{\text{fid}})_i \,, \tag{23}$$

$$\boldsymbol{\Sigma}(\theta_{\text{fid}}) = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{d}(\theta_{\text{fid}})_i - \boldsymbol{\mu}(\theta_{\text{fid}})) \, (\boldsymbol{d}(\theta_{\text{fid}})_i - \boldsymbol{\mu}(\theta_{\text{fid}}))^T \,, \tag{24}$$

$$\left. \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}_m} \right|_{\theta_{\text{fid}}} = \frac{1}{M} \sum_{i=1}^{M} \frac{\boldsymbol{d}(\theta_{\text{fid}} + \delta\theta_m/2)_i - \boldsymbol{d}(\theta_{\text{fid}} - \delta\theta_m/2)_i}{\delta\theta_m} \,, \tag{25}$$

$L_{\mathcal{F}}(\boldsymbol{\Sigma}(\theta_{\text{fid}}))$ is the Frobenius norm between the covariance and the unit matrix $\mathbb{1}$, $L_2(W_{NN})$ is the l_2 norm of the networks' weights. The final loss is then the sum of the negative logarithm of the Fisher information, together with the covariance and weight regularizers. The tanh term in the covariance regularizer is used to turn-off the regularization once the covariance becomes close to the unity.

The main goal of the IMNN is therefore to maximize the Fisher information, while keeping the covariance of the summary fixed to unity. The reason we have the flexibility to normalize the covariance of the summary is the following. For any summary $\boldsymbol{d}$ of dimensionality D, one can define a new summary, $\boldsymbol{d'} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{d}$, where $\boldsymbol{a}$ is a constant D-dim vector serving to normalize $\boldsymbol{d'}$ to be zero mean, and $\boldsymbol{B}$ is a nonsingular $D \times D$ matrix, serving to normalize $\boldsymbol{d'}$ to be unit variance ($\boldsymbol{B}$ is equivalent to a zero-phase component analysis (ZCA) whitening kernel; see Kessy et al. 2018). One can then show that $\ln P(\boldsymbol{d}|\theta) = \ln P(\boldsymbol{d'}|\theta) + \ln \det \boldsymbol{B}$, or equivalently $\partial \ln P(\boldsymbol{d}|\theta)/\partial\theta_m = \partial \ln P(\boldsymbol{d'}|\theta)/\partial\theta_m$. This means there is a freedom to transform the summary $\boldsymbol{d}$ to be zero-mean, unit-variance at a single point θ_{fid}, without changing the Fisher information anywhere in the parameter space.

We also note that the Fisher matrix used for computing the IMNN loss is calculated with two approximations: (i) using only the first term; and (ii) without the inverse covariance between the two derivatives (see Eq. (4)). Both of these simplifications were needed in order to stabilize the training of the IMNN, and they are correct as long as the covariance is kept at unity[11].

For the IMNN training, we created a database around the fiducial θ_{fid} (see Figs. 1 and 5) for $N = M = 1024$ (Eqs. (23), (24), (25)), with 11 264 realizations in total. The training and validation sets were formed by splitting the database in half, following the original work of Charnock et al. (2018). We trained the network using stochastic gradient descent with a batch size $b = 128$, regularization strengths $\lambda = 10$, $\lambda_W = 10^{-9}$ and the Adam optimizer. To increase the effective database size and improve the training, the telescope noise was added to the simulations "on the fly". The final model has been trained on 8 V-100 GPUs in parallel, for $\approx$120k epochs and $\approx$ 20k GPUh. The IMNN code developed for the purposes of this work we make publicly available under 21cmIMNN[12] package.

3.6. 2DPS + IMNN

Finally, we also computed the Fisher information of the concatenation of the individual 2DPS and IMNN summaries. As we shall see in the following section, these are the two best performing individual summaries. We note that one can cleanly combine summaries when performing simulation-based inference (SBI), as the likelihood does not need to be specified explicitly. The main restriction is the size of the required database and the stability of the SBI training (e.g., Cranmer et al. 2020; Lueckmann et al. 2021). When computing the Fisher information in this work, we were restricted to these two summaries due to the limits imposed by our database size. In particular, the minimal number of samples needed to estimate the covariance matrix is proportional to the dimensionality of the total summary (e.g., Hartlap et al. 2007). Furthermore, the dimensions of the final summary should be uncorrelated, as otherwise the covariance is singular and non-invertible. This can result in incorrect Fisher matrix estimation (see however Tegmark et al. 1997 for possible alternative solutions).

4. Database

Since we did not have a strongly motivated fiducial choice for our galaxy parameters, we wanted to compare the Fisher information of 21 cm summaries across our prior volume. To do so, we constructed a "Fisher database", sampling 152 parameter combinations from our prior. We used a flat prior for the X-ray emission parameters, L_X and E_0, over the ranges shown in Fig. 5 (for more motivation of these choices see Fragos et al. 2013; Lehmer et al. 2016; Das et al. 2017). The remaining parameters characterizing star formation and ionizing emission of galaxies are constrained by observed UV luminosity functions (e.g., Bouwens et al. 2015, 2017; Oesch et al. 2018) as well as estimates of the reionization history (e.g., McGreer et al. 2015; Planck Collaboration Int. XLVII 2016). As our model is nested in the 6D Park et al. (2019) model, we used the posterior from that work, calculating the corresponding conditional distributions over $f_{*,10}$, $f_{\text{esc},10}$ and M_{turn}, fixing the remaining parameters to their fiducial values, as discussed in the previous section. To compute the conditional distributions, we used the conditional_kde[13] code.

Our prior and samples are shown in Fig. 5. The diagonal panels illustrate the 1D prior probability distribution functions (PDFs), while the contours in the lower left enclose 95% of the 2D distributions. The 152 samples are shown as points, while

---

[11] The original work by Charnock et al. (2018) had slightly different loss function, by having the inverse of the covariance in the Fisher term and the additional Frobenius norm between $\boldsymbol{\Sigma}^{-1}$ and $\mathbb{1}$. We find that these choices destabilized the training and are not crucial for the convergence (see also Makinen et al. 2021).

[12] https://github.com/dprelogo/21cmIMNN
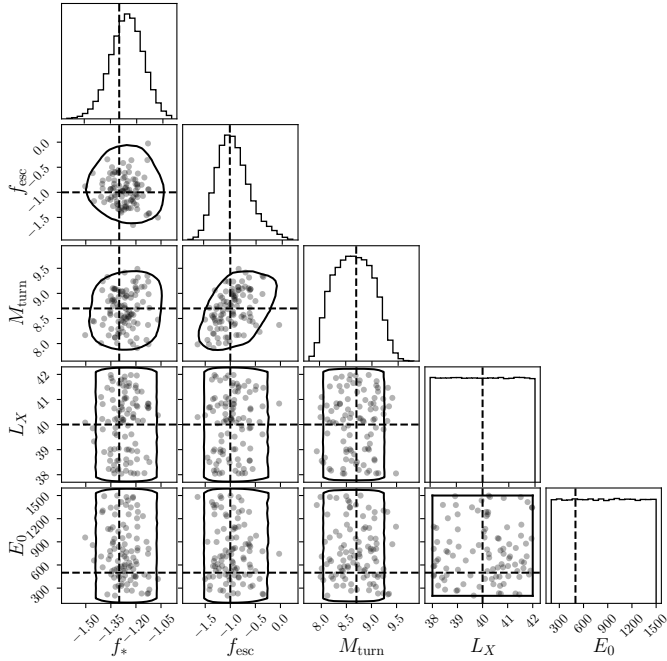[13] https://github.com/dprelogo/conditional_kde

Fig. 5. Prior volume of the astrophysical parameter space. Dashed lines demarcate the fiducial parameter values used to train the IMNN, 2D contours enclose 95% of the prior volume, and the points denote the 152 prior samples at which we computed the Fisher information for our summaries. Note that all units and $\log_{10}$ factors have been removed for clarity; see the main text and Fig. 1 for reference on the units.

the dashed vertical and horizontal lines demarcate the fiducial choice used to train our IMNN.

We simulated $N = 130$ realizations for each of the 152 parameter samples, $\boldsymbol{\theta}^*$, in order to estimate the mean and the covariance of each summary (Eqs. (5), (6)), plus an additional $2 \times 5 \times M = 150$ realizations at $\boldsymbol{\theta}^* \pm \delta\theta_m/2$ to estimate the derivative of the mean (Eq. (10)) along each of the 5 parameter basis vectors. We note that no new simulations are needed to estimate the derivative of the covariance matrix, since we used the same $L = M = 15$ realizations that were used to compute the first Fisher term mentioned above (cf. Eq. (11) and associated discussion). In Appendix D we show convergence tests of these choices for estimating the Fisher information of our summaries. We caution that the WST and RNN summaries require a larger dataset to fully converge.

5. Results

We now present the main results of this work. We first discuss the training of the IMNN in Section 5.1, then compare the performance of all of our 21 cm summaries in Section 5.2, adding a final discussion on the importance of including the second Fisher term (cf. Eq. (4)) in Section 5.3. Throughout, our figure of merit will be the PDF of the Fisher information ($\log_{10}(\det \boldsymbol{F})$), over the prior volume (see Sect. 4). We quote both the median and the variance of this distribution. A high value of the median means that the typical parameter value is well constrained by the summary, while a narrow variance means that the constraining power of the summary is relatively constant over the prior volume. We note that the Fisher information is a measure of the average constraining power over our five parameters; we do not discuss the relative constraints on individual parameters in this work.
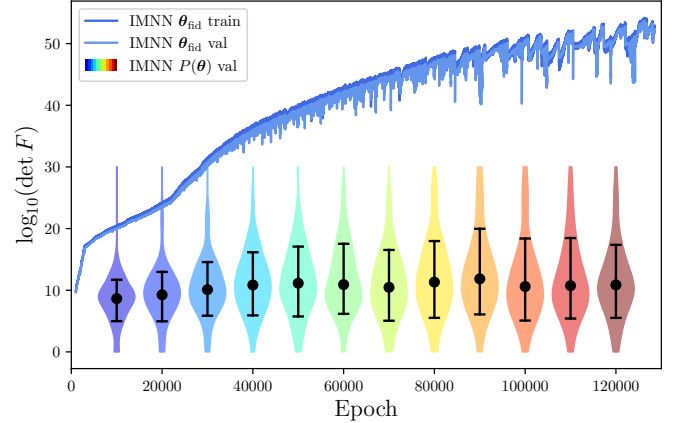
Fig. 6. Fisher information of the IMNN during training. Dark (light) blue curves show the Fisher information for the training (validation) set of the IMNN trained at $\boldsymbol{\theta}_{\mathrm{fid}}$. The violin plots show the distribution of the Fisher information over the full prior volume excluding $\boldsymbol{\theta}_{\mathrm{fid}}$, $P(\theta)$, at every interval of the 10k training epochs. Black dots and error bars inside each violin denote the median and 68% CLs of the distributions. The divergence between the Fisher information computed at $\boldsymbol{\theta}_{\mathrm{fid}}$ and its PDF over the remaining prior volume clearly indicates that the IMNN is over-specializing to the specific parameter combination, $\boldsymbol{\theta}_{\mathrm{fid}}$, used in the training.

5.1. IMNN training

In Fig. 6 we show the training results for the IMNN. Blue lines show the training and validation Fisher information (see the first term of the IMNN loss, Eq. (22)) at the fiducial point $\boldsymbol{\theta}_{\mathrm{fid}}$. We recall that the training and validation sets each consist of 5632 light cone realizations, evaluated at $\boldsymbol{\theta}_{\mathrm{fid}}$ with different cosmic initial conditions and telescope noise seeds. The Fisher information of the training and validation sets are largely overlapping meaning the network does not overfit to the training set. We see that by the end of the training, the IMNN has learned to improve the Fisher information by 40(!) orders of magnitude, corresponding to parameter constraints that are improved by an average factor of $\sim 10^8$ in the variance!

Is the compression learned by the IMNN comparably informative at other parameter combinations? We can answer this question by comparing the solid lines of Fig. 6 to the violin distributions. The latter show the distribution of the Fisher information over the full prior volume, $P(\theta)$, at every interval of 10k training epochs, excluding $\boldsymbol{\theta}_{\mathrm{fid}}$. We see that although the IMNN is not overfitting in the classical sense (the Fisher information of the validation set keeps increasing), it quickly "over-specializes" to the specific parameter value used for training, $\boldsymbol{\theta}_{\mathrm{fid}}$. Thus, the compression learned by the IMNN does not generalize well across parameter space[14]. The PDF of the Fisher information over the prior volume does not change significantly after epoch 40k, with the median beginning to decrease beyond epoch 90k. We took epoch 90k as the final state of the IMNN in the remainder of this work.

---

[14] One solution could be to train the IMNN around several different points. However, this is extremely expensive in our application, both in terms of database size and GPU time. Furthermore, it is not clear if such a procedure would result in a meaningful summary, as the concept of one point training is inherently built into the IMNN. Alternatively, it might be possible to construct a generalized IMNN, maximizing the Fisher information throughout parameter space; however, it is not obvious how this could be done.
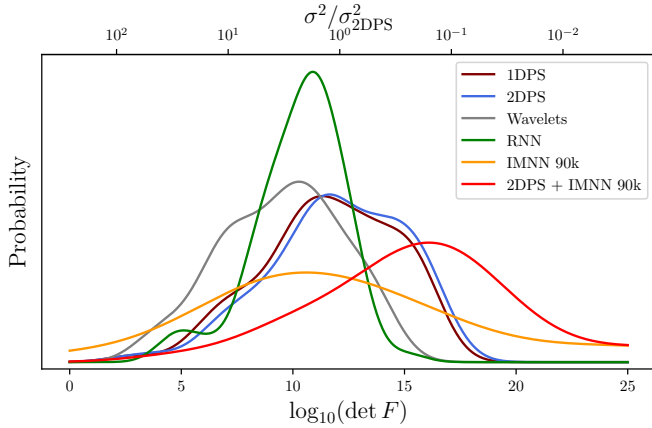
Fig. 7. Distribution of the Fisher information (det F) across the prior volume, for all summaries considered in this work. For the IMNN summary, we used the weights that give the highest median value (corresponding to 90k epochs; see Fig. 6). On the top axis we denote the corresponding relative improvement in the single parameter variance (Eq. (26)), normalized to its median value of the 2DPS summary. The 2DPS has the largest median value of all of the individual summaries, while the combination of the 2DPS + IMNN dramatically outperforms every individual summary. See Table 1 for quantitative values.

5.2. Comparison of all summaries

In Fig. 7 we show the main result of this work: the distribution of Fisher information across the prior volume, for all of the summary statistics we consider. On the top axis we denote the corresponding relative improvement in constraints on the single parameter variance σ^2. We define it as an effective variance in each dimension with no correlation, giving the same Fisher information. For a 5D parameter space we have

$$\sigma^2 = (\det F)^{-1/5} . \tag{26}$$

Here, one can ignore the difference in the units (i.e., scaling) between parameter dimensions (see Appendix C for more details). Specifically, the top axis of Fig. 7 denotes the single parameter variance normalized to its median value of the 2DPS summary. To give an example, a factor of 10 times larger Fisher information for a 5-dim parameter space would result in a 37% smaller single parameter variance. The median values and 68% confidence levels (CLs) of the Fisher information and single parameter variance are also listed in Table 1.

Among the individual summaries, the 2DPS is a clear winner. In comparison, the 1DPS has roughly a factor of 2 smaller Fisher information across the prior volume, which translates to a ~15% larger single parameter variance (Eq. (26)). This difference between the 1DPS and 2DPS shows that indeed the anisotropy of the cosmic 21 cm signal helps in parameter recovery, though not very significantly. The distribution of the 2DPS Fisher information is also relatively narrow, indicating its constraining power does not vary enormously across parameter space, compared to most other summaries.

The median Fisher information of the IMNN is comparable with that of the 1DPS. However, the IMNN results in the widest distribution of all of the summaries. This means that the compression that the IMNN learned at θ_{fid} can be much more informative but also much less informative, depending on the parameter choice. Therefore, if one has a good idea of where the maximum likelihood value will be (e.g., from complementary observations; Park et al. 2019; Abdurashidova et al. 2022;

Table 1. Distributions of the Fisher information.

Summary	Median$^{+34\%}_{-34\%}$	σ^2/σ^2_{2DPS}
Wavelets	$9.6^{+2.9}_{-3.1}$	3.16
RNN	$10.5^{+1.8}_{-2.2}$	2.09
IMNN 90k	$11.8^{+8.2}_{-5.7}$	1.15
1DPS	$11.8^{+3.1}_{-3.4}$	1.15
2DPS	$12.1^{+3.1}_{-3.3}$	1.00
IMNN 90k + 2DPS	$15.9^{+5.2}_{-4.9}$	0.17

Notes. Companion table to Fig. 7, summarizing the Fisher information probability distribution across the prior volume for each summary. In particular, the first column represents values of $\log_{10}(\det F)$ for the median and 68% of probability volume around it. The second column shows the single parameter variance (Eq. (26)), which serves as a guide to how much would the recovered variance of a parameter increase or decrease if one uses that particular summary instead of the 2DPS.

HERA Collaboration 2023; Breitman et al. 2024), it could be beneficial to train the IMNN at that best guess value and use the resulting summary in inference. An optimal strategy could be to perform inference using the 2DPS in order to obtain the maximum likelihood estimate (MLE), train the IMNN at this MLE parameter combination, then perform inference again using the resulting IMNN summary.

The RNN performs worse than the two PS summaries, resulting in a factor of 2 larger single parameter variance compared to the 2DPS. However, its distribution is the most homogeneous across the prior volume (i.e., the narrowest). This could be interpreted as sacrificing the overall information content for robustness throughout the parameter space. The RNN was trained as a regressor over a large parameter space, and so it is understandable that it is much more consistent with respect to the IMNN, which was trained on a single parameter combination. The overall low median information content of the RNN compression could be due to the fact that it was trained on a different parametrization of the cosmological signal than used in this work. We expect that retraining the RNN as a regressor on the same parametrization would improve the median information content. Indeed, Zhao et al. (2022a) find that a regressor CNN slightly outperforms the 1DPS, resulting in tighter parameter constraints when using two different parameter combinations for the mock observation. However, it is likely that this increase in the information content would come at the cost of robustness (i.e., a broader PDF), as the RNN over-specializes to features inherent in that model's parametrization and its training set. Additionally, we caution the Fisher estimate for the RNN summary needs a larger dataset to fully converge (cf. Appendix D).

The wavelet summary results in a median Fisher information that is a factor of ~300 smaller than that of the 2DPS at their median values (factor of ~3 increase in the single parameter variance (Eq. (26)). Intuitively, one would expect more information content in wavelets compared with the PS, as they effectively include higher-order correlations. However, our WST followed the definition in Greig et al. (2022, 2023), in which the wavelets are only computed at one slice for each redshift chunk. Thus, the physics-rich, line-of-sight modes inside each redshift bin are lost for this choice of WST. The weaker performance of the WST (as defined in this work) compared to the power spectra implies that there is more information in the line-of-sight modes than there is in higher-order correlations of the transverse modes.

We note that recent works also using 2D wavelets achieved better results than we do here (Greig et al. 2022; Hothi et al. 2024). However, their analysis differs from ours. In particular, they use more redshift bins and they do not average the coefficients over the sky-plane (rolling the filter as described above). We confirm that without the additional smoothing that we performed in this work, the Gaussian approximation for the likelihood that is intrinsic to the Fisher estimate (cf. Eq. (4)) is notably worse, for both the wavelets and the RNN. Therefore, the higher wavelet information found by Greig et al. (2022), Hothi et al. (2024) could be partially due to a less appropriate application of the Fisher estimate. Alternatively, the information content could be improved by using 3D wavelets instead of 2D ones (see, e.g., Zhao et al. 2023a). As for the RNN, we caution that the WST needs a larger dataset for the Fisher to fully converge (cf. Appendix D).

Finally, our concatenation of the 2DPS and IMNN significantly outperforms all of the individual summaries (cf. the red curve in Fig. 7). As we saw above, these two summary statistics are highly complimentary: the 2DPS is robust (i.e., a narrow information PDF), while the IMNN summary can be much more constraining for some parameters but much less for others (i.e., it has a very broad information PDF). The distribution of Fisher information of 2DPS + IMNN retains the narrowness of the 2DPS, but is shifted to the right, including the high-value tail from the IMNN[15]. The combined summary is a factor of ~6.5–9.5 better in the single parameter variance (Eq. (26)), averaged over the prior samples. Combining multiple 21 cm statistics has already been proven to tighten parameter constraints (e.g., Gazagnes et al. 2021; Watkinson et al. 2022). However, previous work using traditional inference was limited by the need to derive a tractable likelihood, which becomes impossible for nontrivial summaries. Fortunately, recent advances in applying SBI to 21 cm (e.g., Zhao et al. 2022a,b; Prelogović & Mesinger 2023; Saxena et al. 2023) will allow us to cleanly combine multiple summary statistics without having to explicitly define the likelihood. Unfortunately, the neural density estimation that is intrinsic to SBI can be unstable and/or require large training sets, if the summary statistics become very high dimensional. This work illustrates how combining only two complimentary statistics can result in much tighter, more robust parameter constraints than would be available using either individually.

5.3. One-term versus full Fisher matrix

The full Fisher matrix of a Gaussian likelihood can be separated in two terms $F = F_\mu + F_\Sigma$ (see Eq. (4)), where the first term includes the gradient of the mean and the second term the gradient of the covariance matrix (e.g., Appendix B; Tegmark et al. 1997). All of our results from the previous sections were calculated using the full expression, including both terms. However, it is common in the literature to ignore one of the terms, by approximating the mean or the covariance matrix as a constant (e.g., Carron 2013). In most of cosmological Fisher forecasts using the PS (e.g., Sailer et al. 2021; Abazajian et al. 2022; Euclid Collaboration 2022; Mason et al.

Fig. 8. Distributions of Fisher information for the 1DPS (red) and 2DPS (blue). Solid lines account for both terms in the Fisher matrix (cf. Eq. (4)), while the dashed lines correspond to the common simplification of a constant covariance (i.e., including only the first term in Eq. (4)). The assumption of a constant covariance underpredicts the Fisher information by a factor of ~2. For a 5-parameter model, this translates to a ~15% larger single parameter variance (Eq. (26)).

2023; Bykov et al. 2023), the second term, F_Σ, is neglected[16]. In some cases, the covariance matrix can indeed be independent or only weakly dependent on the parameters, justifying ignoring the second Fisher term. However, the importance of the F_Σ term for 21 cm Fisher forecasts has not yet been investigated (e.g., Greig et al. 2022; Balu et al. 2023; Mason et al. 2023; Hothi et al. 2024). Indeed in Prelogović & Mesinger (2023) we showed that assuming a parameter-independent covariance matrix of the 21 cm 1DPS likelihood can bias the posterior estimation. In this section, we quantify the importance of the F_Σ term in determining the Fisher information[17], for all of our 21 cm summaries. It is important to note that adding this term will always increase the information content, as $\det(F_\mu + F_\Sigma) \geq \det F_\mu$. More details are given at the end of Appendix A.

In Fig. 8 we plot the Fisher information PDF of the commonly used 1DPS and 2DPS. The solid curves are the same as in the previous figure, corresponding to the full expression for the Fisher information. The dashed curves only include the first term, assuming a constant covariance matrix. For both 1D and 2D power spectra, neglecting the second term underestimates the Fisher information by a factor of ~2.3 on average over the prior samples. For our 5-parameter model, this translates into an underestimate of the single parameter variance (Eq. (26)) by 15%.

In Fig. 9 we compare the Fisher information PDFs with and without the second Fisher term for the remaining individual summaries. We see that for the other 21 cm summaries, the second Fisher term is even more important than for the power spectra. For example, assuming a constant covariance underestimates the mean Fisher information of the RNN for a factor of ~2000 and a factor of ~11 100 for the wavelets, averaged over the prior samples. We note that although the IMNN is trained to normalize the covariance to unity, thus ensuring a constant covariance matrix by construction, this is only done at $\theta_{\rm fid}$. For the other points in the parameter space, there is no guarantee that this condition is met.

---

[15] We note that in the case in which some value of the IMNN summary would be completely degenerate with some 2DPS bin, the resulting covariance matrix would be singular and therefore non-invertible. This would make estimating the Fisher matrix unreliable. However, as the IMNN compression and 2DPS are very different summaries, this is unlikely to be a problem.
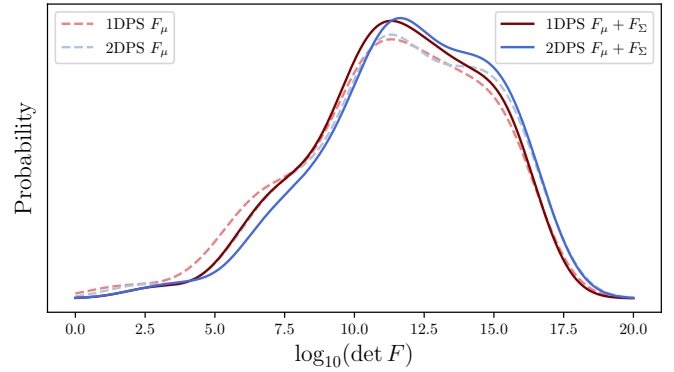
[16] For examples of ignoring the first term, see, e.g., Abramo (2012) and d'Assignies et al. (2023).

[17] Assuming a constant covariance matrix can result in other biases at the level of the full Fisher matrix. For example, it could impact parameter degeneracies and/or improve constraints of some parameters at the cost of others. Here we only focus on the effect this simplification has on the determinant of the Fisher matrix as this quantity measures the average parameter constraining power of the summary statistics.
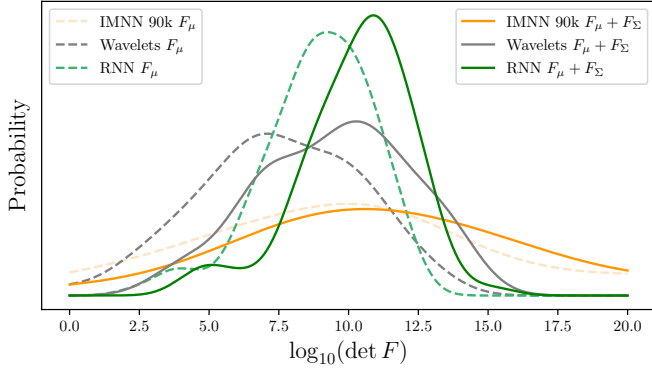
Fig. 9. Analogous to Fig. 8, but for the other 21 cm summary statistics.

6. Conclusions

The cosmic 21 cm signal will revolutionize our understanding of the CD and EoR. "Big Data" analysis techniques will be required to cope with radio telescopes such as the SKA. In this work we compare several approaches to compressing 21 cm light cone data, on the basis of their Fisher information. Specifically, we compare the determinant of the Fisher information for the following summary statistics: (i) the 1DPS, (ii) the 2DPS, (iii) the 2D WST, (iv) a RNN trained as a regressor; (v) an IMNN; and (vi) the combination of 2DPS and IMNN. Importantly we compare their Fisher information across the parameter space since there is currently no standard model for the astrophysics of the first galaxies, which determines the cosmic 21 cm signal.

The 2DPS is the individual summary with the highest median Fisher information (Fig. 7, Table 1). The distribution of its Fisher information is also relatively narrow across the parameter space. This means that compared to most other summaries, the 2DPS can recover astrophysical parameters with relatively similar precision regardless of what the "true" values of our Universe are. We also show that the 2DPS slightly outperforms the 1DPS, confirming that the anisotropy of the 21 cm signal helps constrain parameters. The information content of both the RNN and wavelets is generally lower than either the 1DPS or the 2DPS. However, we caution that both of these summaries could be defined in several different ways, potentially improving their performances (e.g., Greig et al. 2022, 2023; Hothi et al. 2024; Zhao et al. 2023a). Moreover, our Fisher estimates for the RNN and wavelet summaries require a larger dataset to fully converge (see Appendix D).

In this work we also introduce IMNNs to the field of 21 cm cosmology. Although capable of achieving the highest Fisher information for some parameter choices, the IMNN does not generalize well, resulting in a broad distribution across the prior volume. We find an enormous difference between its Fisher information computed at the parameter set used for training, $\boldsymbol{\theta}_{\mathrm{fid}}$, and its information computed at the remaining prior samples (Fig. 6). This means that the IMNN overfits to the information content of the fiducial. The regularization inherent to the IMNN loss (see Eq. (22) and the following discussion) makes it difficult to avoid this overfitting by training over many different choices of $\boldsymbol{\theta}_{\mathrm{fid}}$. A simpler solution could be to first perform inference with a less optimal summary, in order to determine a maximum likelihood point, and then train the IMNN at that point.

Combining the IMNN with the 2DPS results in a summary whose median Fisher information is almost ~4 orders of magnitude larger than the 2DPS alone. This results in a factor of ~6.5–9.5 tighter single parameter variance averaged over the prior samples. This means the IMNN extracts complementary information with respect to the 2DPS. Performing SBI using such a summary could yield extremely tight and robust parameter constraints. Moreover, combining only these two complementary summaries still preserves a high level of compression, with our light cone of ~1.6×10^6 cells reduced to just 77 numbers.

Finally, we stress the importance of not assuming a constant covariance when performing Fisher forecasts using 21 cm observables. This common assumption underestimates the Fisher information of the PS on average by a factor of ~2.3, corresponding to a 15% increase in the single parameter variance (Eq. (26)). For other summaries, the assumption underestimates the Fisher information by orders of magnitude.

References

Abazajian, K., Addison, G. E., Adshead, P., et al. 2022, ApJ, 926, 54
Abdurashidova, Z., Aguirre, J. E., Alexander, P., et al. 2022, ApJ, 924, 51
Abramo, L. R. 2012, MNRAS, 420, 2042
Allys, E., Levrier, F., Zhang, S., et al. 2019, A&A, 629, A115
Alvarez, I., Niemi, J., & Simpson, M. 2014, arXiv e-prints [arXiv:1408.4050]
Anden, J., & Mallat, S. 2014, IEEE Trans. Signal Proc., 62, 4114
Balu, S., Greig, B., & Wyithe, J. S. B. 2023, MNRAS, 525, 3032
Barkana, R., & Loeb, A. 2005, ApJ, 624, L65
Barkana, R., & Loeb, A. 2006, MNRAS, 372, L43
Barnard, J., McCulloch, R., & Meng, X.-L. 2000, Stat. Sin., 10, 1281
Bayer, A. E., Modi, C., & Ferraro, S. 2023, JCAP, 2023, 046
Bharadwaj, S., & Ali, S. S. 2005, MNRAS, 356, 1519
Bharadwaj, S., & Pandey, S. K. 2005, MNRAS, 358, 968
Bouwens, R. J., Illingworth, G. D., Oesch, P. A., et al. 2015, ApJ, 803, 34
Bouwens, R. J., Oesch, P. A., Illingworth, G. D., Ellis, R. S., & Stefanon, M. 2017, ApJ, 843, 129
Breitman, D., Mesinger, A., Murray, S. G., et al. 2024, MNRAS, 527, 9833
Bruna, J., & Mallat, S. 2012, arXiv e-prints [arXiv:1203.1513]
Brunel, N., & Nadal, J.-P. 1998, Neural Comput., 10, 1731
Bykov, S., Gilfanov, M., & Sunyaev, R. 2023, A&A, 669, A61
Carron, J. 2013, A&A, 551, A88
Charnock, T., Lavaux, G., & Wandelt, B. D. 2018, Phys. Rev. D, 97, 083004
Chen, Z., Xu, Y., Wang, Y., & Chen, X. 2019, ApJ, 885, 23
Cheng, S., Ting, Y.-S., Ménard, B., & Bruna, J. 2020, MNRAS, 499, 5902
Cooray, A., Li, C., & Melchiorri, A. 2008, Phys. Rev. D, 77, 103506
Cramér, H. 1999, Mathematical Methods of Statistics (Princeton University Press), 43
Cranmer, K., Brehmer, J., & Louppe, G. 2020, Proc. Nat. Academy Sci., 117, 30055
Dai, B., & Seljak, U. 2022, MNRAS, 516, 2363
Das, A., Mesinger, A., Pallottini, A., Ferrara, A., & Wise, J. H. 2017, MNRAS, 469, 1166
d'Assignies, D. W., Zhao, C., Yu, J., & Kneib, J. P. 2023, MNRAS, 521, 3648
Datta, K. K., Mellema, G., Mao, Y., et al. 2012, MNRAS, 424, 1877
Doussot, A., & Semelin, B. 2022, A&A, 667, A118
Euclid Collaboration (Ilić, S., et al.) 2022, A&A, 657, A91
Fefferman, C., Mitter, S., & Narayanan, H. 2016, J. Am. Math. Soc., 29, 983
Fisher, R. A. 1935, J. R. Stat. Soc., 98, 39
Flöss, T., de Wild, T., Meerburg, P. D., & Koopmans, L. V. E. 2022, JCAP, 2022, 020
Fragos, T., Lehmer, B., Tremmel, M., et al. 2013, ApJ, 764, 41

Gabor, D. 1946, J. Inst. Electrical Eng. III: Radio Commun. Eng., 93, 429
Gazagnes, S., Koopmans, L. V. E., & Wilkinson, M. H. F. 2021, MNRAS, 502, 1816
Gehlot, B. K., Mertens, F. G., Koopmans, L. V. E., et al. 2020, MNRAS, 499, 4158
Gillet, N., Mesinger, A., Greig, B., Liu, A., & Ucci, G. 2019, MNRAS, 484, 282
Giri, S. K., & Mellema, G. 2021, MNRAS, 505, 1863
Giri, S. K., Mellema, G., Dixon, K. L., & Iliev, I. T. 2018, MNRAS, 473, 2949
Gleser, L., Nusser, A., Ciardi, B., & Desjacques, V. 2006, MNRAS, 370, 1329
Goupillaud, P., Grossmann, A., & Morlet, J. 1984, Geoexploration, 23, 85
Greig, B., & Mesinger, A. 2015, MNRAS, 449, 4246
Greig, B., & Mesinger, A. 2018, MNRAS, 477, 3217
Greig, B., Ting, Y.-S., & Kaurov, A. A. 2022, MNRAS, 513, 1719
Greig, B., Ting, Y.-S., & Kaurov, A. A. 2023, MNRAS, 519, 5288
Harker, G. J. A., Zaroubi, S., Thomas, R. M., et al. 2009, MNRAS, 393, 1449
Hartlap, J., Simon, P., & Schneider, P. 2007, A&A, 464, 399
Heavens, A. F., Jimenez, R., & Lahav, O. 2000, MNRAS, 317, 965
Heneka, C. 2023, Deep Learning 21 cm Lightcones in 3D, in Machine Learning for Astrophysics, ML4Astro 2022, eds. F. Bufano, S. Riggi, E. Sciacca, F. Schilliro (Springer)
HERA Collaboration (Abdurashidova, Z., et al.) 2023, ApJ, 945, 124
Hochreiter, S., & Schmidhuber, J. 1997, Neural Comput., 9, 1735
Hong, S. E., Ahn, K., Park, C., et al. 2014, J. Korean Astron. Soc., 47, 49
Hothi, I., Allys, E., Semelin, B., & Boulanger, F. 2024, A&A, 686, A212
Jasche, J., & Kitaura, F. S. 2010, MNRAS, 407, 29
Jasche, J., & Wandelt, B. D. 2012, MNRAS, 425, 1042
Jasche, J., & Wandelt, B. D. 2013, MNRAS, 432, 894
Jiang, J., Shu, Y., Wang, J., & Long, M. 2022, arXiv e-prints [arXiv:2201.05867]
Kapahtia, A., Chingangbam, P., & Appleby, S. 2019, JCAP, 2019, 053
Kapahtia, A., Chingangbam, P., Ghara, R., Appleby, S., & Choudhury, T. R. 2021, JCAP, 2021, 026
Kessy, A., Lewin, A., & Strimmer, K. 2018, Am. Stat., 72, 309
Kitaura, F. S., & Enßlin, T. A. 2008, MNRAS, 389, 497
Kittiwisit, P., Bowman, J. D., Murray, S. G., et al. 2022, MNRAS, 517, 2138
Koopmans, L., Pritchard, J., Mellema, G., et al. 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14), 1
Kwon, Y., Hong, S. E., & Park, I. 2020, J. Korean Phys. Soc., 77, 49
La Plante, P., & Ntampaka, M. 2019, ApJ, 880, 110
Leclercq, F., Jasche, J., Lavaux, G., Wandelt, B., & Percival, W. 2017, JCAP, 2017, 049
Lehmer, B. D., Basu-Zych, A. R., Mineo, S., et al. 2016, ApJ, 825, 7
Lin, Y., Oh, S. P., Furlanetto, S. R., & Sutter, P. M. 2016, MNRAS, 461, 3361
Liu, A., Parsons, A. R., & Trott, C. M. 2014a, Phys. Rev. D, 90, 023018
Liu, A., Parsons, A. R., & Trott, C. M. 2014b, Phys. Rev. D, 90, 023019
Livet, F., Charnock, T., Le Borgne, D., & de Lapparent, V. 2021, A&A, 652, A62
Loeb, A., & Zaldarriaga, M. 2004, Phys. Rev. Lett., 92, 211301
Lu, T.-Y., Mason, C. A., Hutter, A., et al. 2024, MNRAS, 528, 4872
Lueckmann, J.-M., Boelts, J., Greenberg, D. S., Gonçalves, P. J., & Macke, J. H. 2021, arXiv e-prints [arXiv:2101.04653]
Majumdar, S., Pritchard, J. R., Mondal, R., et al. 2018, MNRAS, 476, 4007
Makinen, T. L., Charnock, T., Alsing, J., & Wandelt, B. D. 2021, JCAP, 2021, 049
Makinen, T. L., Charnock, T., Lemos, P., et al. 2022, Open J. Astrophys., 5, 18
Makinen, T. L., Alsing, J., & Wandelt, B. D. 2023, JMLR submitted, [arXiv:2310.03812]
Mallat, S. 2011, arXiv e-prints [arXiv:1101.2286]
Mangena, T., Hassan, S., & Santos, M. G. 2020, MNRAS, 494, 600
Mao, Y., Shapiro, P. R., Mellema, G., et al. 2012, MNRAS, 422, 926
Mason, C. A., Muñoz, J. B., Greig, B., Mesinger, A., & Park, J. 2023, MNRAS, 524, 4711

McAlpine, S., Helly, J. C., Schaller, M., et al. 2022, MNRAS, 512, 5823
McGreer, I. D., Mesinger, A., & D'Odorico, V. 2015, MNRAS, 447, 499
McQuinn, M., & O'Leary, R. M. 2012, ApJ, 760, 3
Mellema, G., Iliev, I. T., Pen, U.-L., & Shapiro, P. R. 2006, MNRAS, 372, 679
Mertens, F. G., Mevius, M., Koopmans, L. V. E., et al. 2020, MNRAS, 493, 1662
Mesinger, A. 2020, The Cosmic 21-cm revolution: charting the First Billion Years of Our Universe (IOP Publishing), 2514
Mesinger, A., & Furlanetto, S. 2007, ApJ, 669, 663
Mesinger, A., Furlanetto, S., & Cen, R. 2011, MNRAS, 411, 955
Mesinger, A., Ewall-Wice, A., & Hewitt, J. 2014, MNRAS, 439, 3262
Mondal, R., Bharadwaj, S., & Datta, K. K. 2018, MNRAS, 474, 1390
Morales, M. F., Hazelton, B., Sullivan, I., & Beardsley, A. 2012, ApJ, 752, 137
Munshi, S., Mertens, F. G., Koopmans, L. V. E., et al. 2024, A&A, 681, A62
Murray, S. G., & Trott, C. M. 2018, ApJ, 869, 25
Murray, S., Greig, B., Mesinger, A., et al. 2020, J. Open Source Softw., 5, 2582
Oesch, P. A., Bouwens, R. J., Illingworth, G. D., Labbé, I., & Stefanon, M. 2018, ApJ, 855, 105
Pacucci, F., Mesinger, A., Mineo, S., & Ferrara, A. 2014, MNRAS, 443, 678
Park, J., Mesinger, A., Greig, B., & Gillet, N. 2019, MNRAS, 484, 933
Parsons, A. R., Liu, A., Aguirre, J. E., et al. 2014, ApJ, 788, 106
Planck Collaboration VI. 2020, A&A, 641, A6
Planck Collaboration Int. XLVII. 2016, A&A, 596, A108
Prelogović, D., & Mesinger, A. 2023, MNRAS, 524, 4239
Prelogović, D., Mesinger, A., Murray, S., Fiameni, G., & Gillet, N. 2022, MNRAS, 509, 3852
Pritchard, J. R., & Furlanetto, S. R. 2007, MNRAS, 376, 1680
Rao, C. R. 1992, in Breakthroughs in Statistics: Foundations and Basic Theory (Springer), 235
Sailer, N., Castorina, E., Ferraro, S., & White, M. 2021, JCAP, 2021, 049
Saxena, A., Cole, A., Gazagnes, S., et al. 2023, MNRAS, 525, 6097
Schmidt, R. M. 2019, arXiv e-prints [arXiv:1912.05911]
Shi, X., Chen, Z., Wang, H., et al. 2015, arXiv e-prints [arXiv:1506.04214]
Shimabukuro, H., Yoshiura, S., Takahashi, K., Yokoyama, S., & Ichiki, K. 2016, MNRAS, 458, 3003
Shimabukuro, H., Mao, Y., & Tan, J. 2022, Res. Astron. Astrophys., 22, 035027
Sifre, L., & Mallat, S. 2013, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1233
Spall, J. C. 2005, J. Comput. Graphical Stat., 14, 889
Spina, B., Porciani, C., & Schimd, C. 2021, MNRAS, 505, 3492
Sui, C., Zhao, X., Jing, T., & Mao, Y. 2023, arXiv e-prints [arXiv:2307.04994]
Tegmark, M., Taylor, A. N., & Heavens, A. F. 1997, ApJ, 480, 22
Trott, C. M., Wayth, R. B., & Tingay, S. J. 2012, ApJ, 757, 101
Trott, C. M., Jordan, C. H., Midgley, S., et al. 2020, MNRAS, 493, 4711
Vedantham, H., Udaya Shankar, N., & Subrahmanyan, R. 2012, ApJ, 745, 176
Visbal, E., Barkana, R., Fialkov, A., Tseliakhovich, D., & Hirata, C. M. 2012, Nature, 487, 70
Vogeley, M. S., & Szalay, A. S. 1996, ApJ, 465, 34
Watkinson, C. A., & Pritchard, J. R. 2015, MNRAS, 454, 1416
Watkinson, C. A., Giri, S. K., Ross, H. E., et al. 2019, MNRAS, 482, 2653
Watkinson, C. A., Greig, B., & Mesinger, A. 2022, MNRAS, 510, 3838
Yoshiura, S., Shimabukuro, H., Takahashi, K., & Matsubara, T. 2017, MNRAS, 465, 394
Zhao, X., Mao, Y., Cheng, C., & Wandelt, B. D. 2022a, ApJ, 926, 151
Zhao, X., Mao, Y., & Wandelt, B. D. 2022b, ApJ, 933, 236
Zhao, X., Mao, Y., Zuo, S., & Wandelt, B. D. 2023a, ApJ submitted, [arXiv:2310.17602]
Zhao, X., Ting, Y.-S., Diao, K., & Mao, Y. 2023b, MNRAS, 526, 1699

Appendix A: Proof of the determinant inequality

We considered two matrices, $\boldsymbol{C}$ and $\boldsymbol{F}^{-1}$, of shape $m \times m$ for which the following inequality holds:

$$\boldsymbol{C} \geq \boldsymbol{F}^{-1} \, . \tag{A.1}$$

This means that their difference is a positive, semi-definite matrix:

$$\boldsymbol{x}^T (\boldsymbol{C} - \boldsymbol{F}^{-1}) \boldsymbol{x} \geq 0 \quad \forall \boldsymbol{x} \in \mathbb{R}^m \, . \tag{A.2}$$

From this, it follows that

$$\boldsymbol{x}^T \boldsymbol{C} \boldsymbol{x} \geq \boldsymbol{x}^T \boldsymbol{F}^{-1} \boldsymbol{x} \tag{A.3}$$

$$\exp\left(-\frac{1}{2} \boldsymbol{x}^T \boldsymbol{C} \boldsymbol{x}\right) \leq \exp\left(-\frac{1}{2} \boldsymbol{x}^T \boldsymbol{F}^{-1} \boldsymbol{x}\right) \tag{A.4}$$

$$\int d\boldsymbol{x} \, \exp\left(-\frac{1}{2} \boldsymbol{x}^T \boldsymbol{C} \boldsymbol{x}\right) \leq \int d\boldsymbol{x} \, \exp\left(-\frac{1}{2} \boldsymbol{x}^T \boldsymbol{F}^{-1} \boldsymbol{x}\right) \tag{A.5}$$

$$\frac{1}{\sqrt{\det(\boldsymbol{C}/2\pi)}} \leq \frac{1}{\sqrt{\det(\boldsymbol{F}^{-1}/2\pi)}} \tag{A.6}$$

$$\det \boldsymbol{C} \geq \det \boldsymbol{F}^{-1} \, . \tag{A.7}$$

Here the trick was to push the inequality toward the multivariate Gaussian integral. After integrating, what is left is only the inverse of a normalization constant. We note that if one wants to interpret $\boldsymbol{C}$ as a covariance matrix and $\boldsymbol{F}$ as a Fisher matrix, then it would make more sense to integrate over $\boldsymbol{x}^T \boldsymbol{C}^{-1} \boldsymbol{x}$ and $\boldsymbol{x}^T \boldsymbol{F} \boldsymbol{x}$. However, this is irrelevant for the proof, as it would only result in the change of the inequality direction in each row, ending with $\det \boldsymbol{C}^{-1} \leq \det \boldsymbol{F}$. We usually assume the matrices are nonsingular and thus invertible, which is needed here.

We note that the identical procedure can be used to prove that the one-term Fisher is a lower bound of the two-term Fisher. For $\boldsymbol{F} = \boldsymbol{F}_{\mu} + \boldsymbol{F}_{\Sigma}$ one can write $\boldsymbol{x}^T (\boldsymbol{F}_{\mu} + \boldsymbol{F}_{\Sigma}) \boldsymbol{x} \geq \boldsymbol{x}^T \boldsymbol{F}_{\mu} \boldsymbol{x}$, which by following the same steps translates into $\det(\boldsymbol{F}_{\mu} + \boldsymbol{F}_{\Sigma}) \geq \det \boldsymbol{F}_{\mu}$.

Appendix B: Derivation of a Gaussian Fisher information matrix

To derive Eq. 4 (i.e., the Fisher information of a multivariate Gaussian distribution), we started by writing the Fisher information matrix equation (Eq. 1) in the equivalent form:

$$\boldsymbol{F}(\boldsymbol{\theta}^*)_{mn} = -\mathbb{E}\left[\partial^2_{mn} \ln P(\boldsymbol{d}|\boldsymbol{\theta}) \Big| \boldsymbol{\theta}^*\right] \, , \tag{B.1}$$

where we shorten the notation of derivatives as $\partial/\partial\theta_m \equiv \partial_m$ and $\partial^2/\partial\theta_m \partial\theta_n \equiv \partial^2_{mn}$. This form comes from the fact that

$$-\partial^2_{mn} \ln P(\boldsymbol{d}|\boldsymbol{\theta}) = \partial_m \ln P(\boldsymbol{d}|\boldsymbol{\theta}) \cdot \partial_n \ln P(\boldsymbol{d}|\boldsymbol{\theta}) - \frac{1}{P(\boldsymbol{d}|\boldsymbol{\theta})} \partial^2_{mn} P(\boldsymbol{d}|\boldsymbol{\theta}) \, . \tag{B.2}$$

Taking the expectation value on both sides, together with

$$\mathbb{E}\left[\frac{1}{P(\boldsymbol{d}|\boldsymbol{\theta})} \partial^2_{mn} P(\boldsymbol{d}|\boldsymbol{\theta}) \Big| \boldsymbol{\theta}^*\right] = \partial^2_{mn} \int d\boldsymbol{d} \, P(\boldsymbol{d}|\boldsymbol{\theta}^*) = 0 \, , \tag{B.3}$$

proves the equality between Eqs. 1 and B.1.

Before diving into evaluating B.1 for a Gaussian likelihood $\mathcal{N}(\boldsymbol{d}|\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, we firstly list several useful relations:

$$\partial_m \boldsymbol{\Sigma}^{-1} = -\boldsymbol{\Sigma}^{-1} \partial_m \boldsymbol{\Sigma} \, \boldsymbol{\Sigma}^{-1} \, , \tag{B.4}$$

$$\partial_m |\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}| \, \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \partial_m \boldsymbol{\Sigma}\right) \, , \tag{B.5}$$

$$\mathbb{E}\left[A(\boldsymbol{d} - \boldsymbol{\mu})\right] = \boldsymbol{0} \, , \tag{B.6}$$

$$\mathbb{E}\left[(\boldsymbol{d} - \boldsymbol{\mu})^T A (\boldsymbol{d} - \boldsymbol{\mu})\right] = \mathrm{tr}(A\boldsymbol{\Sigma}) \, . \tag{B.7}$$

The first one follows from $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} = \mathbb{1}$ and calculating its derivative, the second one is Jacobi's relation for the derivative of the determinant. The last two are expectation relations for Gaussian distributions for an arbitrary matrix A (where $\dim A = \dim \boldsymbol{\Sigma}$). The first of these follows from the antisymmetry of the integral, and the second is left as an exercise for the conscientious reader.

With this at hand, we can attack the problem of calculating the Fisher matrix for a Gaussian likelihood. By taking the logarithm

$$\ln P(\boldsymbol{d}|\boldsymbol{\theta}) = \mathrm{const.} - \frac{1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} (\boldsymbol{d} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{d} - \boldsymbol{\mu}) \tag{B.8}$$

and computing its second derivative, we get

$$\partial^2_{mn} \ln P(\boldsymbol{d}|\boldsymbol{\theta}) = -\partial_m \boldsymbol{\mu}^T \, \boldsymbol{\Sigma}^{-1} \partial_n \boldsymbol{\mu} \tag{B.9}$$

$$+ \frac{1}{2} \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \partial_m \boldsymbol{\Sigma} \, \boldsymbol{\Sigma}^{-1} \partial_n \boldsymbol{\Sigma}\right) \tag{B.10}$$

$$- (\boldsymbol{d} - \boldsymbol{\mu})^T \, \boldsymbol{\Sigma}^{-1} \partial_m \boldsymbol{\Sigma} \, \boldsymbol{\Sigma}^{-1} \partial_n \boldsymbol{\Sigma} \, \boldsymbol{\Sigma}^{-1} (\boldsymbol{d} - \boldsymbol{\mu}) \tag{B.11}$$

$$- \frac{1}{2} \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \partial^2_{mn} \boldsymbol{\Sigma}\right) \tag{B.12}$$

$$+ \frac{1}{2} (\boldsymbol{d} - \boldsymbol{\mu})^T \, \boldsymbol{\Sigma}^{-1} \partial^2_{mn} \boldsymbol{\Sigma} \, \boldsymbol{\Sigma}^{-1} (\boldsymbol{d} - \boldsymbol{\mu}) \tag{B.13}$$

$$+ \partial^2_{mn} \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{d} - \boldsymbol{\mu}) \tag{B.14}$$

$$- \left(\partial_m \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \partial_n \boldsymbol{\Sigma} + \partial_n \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \partial_m \boldsymbol{\Sigma}\right) \boldsymbol{\Sigma}^{-1} (\boldsymbol{d} - \boldsymbol{\mu}) \, . \tag{B.15}$$

Taking the negative expectation value of the final equation can now be done by using relations B.6 and B.7. One can easily see that only the first three terms contribute, with others canceling each other out or being equal to zero. With this, we finally have

$$\boldsymbol{F}_{mn} = \partial_m \boldsymbol{\mu}^T \, \boldsymbol{\Sigma}^{-1} \partial_n \boldsymbol{\mu} + \frac{1}{2} \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \partial_m \boldsymbol{\Sigma} \, \boldsymbol{\Sigma}^{-1} \partial_n \boldsymbol{\Sigma}\right) \, . \tag{B.16}$$

Appendix C: Fisher information under linear transformations

Imagine we calculate the Fisher information matrix in parameter space $\boldsymbol{\theta}$ around $\boldsymbol{\theta}^*$:

$$\boldsymbol{F}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*)_{mn} = \mathbb{E}\left[\frac{\partial}{\partial\theta_m} \ln P(\boldsymbol{d}|\boldsymbol{\theta}) \cdot \frac{\partial}{\partial\theta_n} \ln P(\boldsymbol{d}|\boldsymbol{\theta}) \Big| \boldsymbol{\theta}^*\right] \, . \tag{C.1}$$

If we now take a linear transformation of the parameter space $\boldsymbol{\eta} = \boldsymbol{a} \odot \boldsymbol{\theta} + \boldsymbol{b}$, where $\odot$ is element-wise multiplication, we get

$$\boldsymbol{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}^* = \boldsymbol{a} \odot \boldsymbol{\theta}^* + \boldsymbol{b}) = A \, \boldsymbol{F}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*) \, A \, , \tag{C.2}$$

where $A = \mathrm{diag}(\boldsymbol{a})$ is a diagonal matrix with elements on the diagonal equal to the vector $\boldsymbol{a}$. That further means:

$$\det \boldsymbol{F}_{\boldsymbol{\eta}} = \det A \boldsymbol{F}_{\boldsymbol{\theta}} A = (\det A)^2 \, \det \boldsymbol{F}_{\boldsymbol{\theta}} \, , \tag{C.3}$$

$$\ln \det \boldsymbol{F}_{\boldsymbol{\eta}} = C + \ln \det \boldsymbol{F}_{\boldsymbol{\theta}} \, . \tag{C.4}$$

The final relation can be useful when thinking about Fisher information calculated in different physical units (different scalings). The only consequence of such a procedure is that the logarithm of the Fisher information shifts by a constant C, independently on the position in the parameter space. This further means that the distribution of the Fisher information shifts by the same constant.

For the single parameter variance definition (Eq. 26), the units (scaling) do not play a role, as we always express it with respect to the 2DPS. Therefore,

$$\frac{\sigma^2}{\sigma^2_{\mathrm{2DPS}}} = \left(\frac{\det \boldsymbol{F}_{\boldsymbol{\theta}}}{\det \boldsymbol{F}_{\boldsymbol{\theta},\mathrm{2DPS}}}\right)^{-1/5} = \left(\frac{\det \boldsymbol{F}_{\boldsymbol{\eta}}}{\det \boldsymbol{F}_{\boldsymbol{\eta},\mathrm{2DPS}}}\right)^{-1/5} \, . \tag{C.5}$$

Appendix D: Convergence tests for the Fisher matrix

When computing the Fisher matrix (Eq. 4), one has to decide on the number of samples from which Σ^{-1}, $\nabla_\theta\mu$ and $\nabla_\theta\Sigma$ are computed. In other words, one has to decide on the size of N, M and L (see Eqs. 5 - 11). In this work, we fixed $M = L$ and used the same samples to compute $\nabla_\theta\mu$ and $\nabla_\theta\Sigma$. Furthermore, we fixed $N = 130$ and $M = 15$. We justified these choices by running a series of convergence tests.

For this test, we used a database of 1024 simulations and their numerical derivatives centered at the fiducial parameters θ_{fid}, used to train the IMNN (see Section 3.5). For that reason, we excluded the IMNN from this test – as the summary space size for IMNN is only 5, it is the least problematic one considering the convergence.

In Figures D.1, D.2, and D.3, we show three slices through the N, M space. In Figure D.1, we fix $N = M$, computing the Fisher information for each summary using five randomly sampled realizations. The solid lines (shaded regions) correspond to the means (standard deviations) computed over these five subsamples. These can be compared to the dashed lines, which show the mean computed over the entire set of 1024 realizations. We see that all of our summaries have converged over the entire range shown except the wavelets, which converge after $N \geq 130$. This result is expected, as wavelets are 112-dim summary, while others are significantly smaller.

In Figure D.2, we fix $M = 15$ and vary N in the same range as in Fig. D.1. Here the black vertical line marks our fiducial choice of $N = 130$. We see that all summaries at $N = 130$ have converged, although the RNN and wavelets show evidence of a small residual bias. The later statistics require a larger M to fully converge to the mean.

Finally, in Figure D.3 we fix $N = 130$ and vary $M \in [5, 200]$, marking $M = 15$ with a black vertical line. Here 1DPS and 2DPS stay almost constant, while RNNs and wavelets show some evolution. Non-monotonic evolution of variance in the RNN might be a consequence of small number of samples on which the lines have been computed. Most importantly, the chosen point of $M = 15$ shows a small bias for all summaries. The quality of the measured Fisher information might be better at $M \approx 100$; however, this would increase the database size by two orders of magnitude, which is computationally unreachable. Importantly, these biases are over an order of magnitude smaller than the summary to summary and parameter to parameter variances computed in the main text; therefore our main conclusions are robust to our database size. However, we caution that RNN and wavelet summaries need a larger M to fully converge.

One could increase the numerical stability and statistical significance of the Fisher estimates by running a Bayesian inference over the covariance matrix and other parameters (see, e.g., Alvarez et al. 2014). This is noted in Hothi et al. (2024), in which the authors study 21 cm Fisher forecasts using different wavelet-based summaries. In our setup, this would mean introducing a prior distribution over the covariance matrix, such as the inverse Wishart distribution (Barnard et al. 2000). One would then estimate the posterior distribution of the covariance from the samples of a summary and marginalize over it while recovering a distribution of the Fisher information. A small step in this direction has been made in this work by properly un-biasing estimates of the covariance and its inverse (Eqs. 7 and 8). A more complete statistical analysis is left for future work.
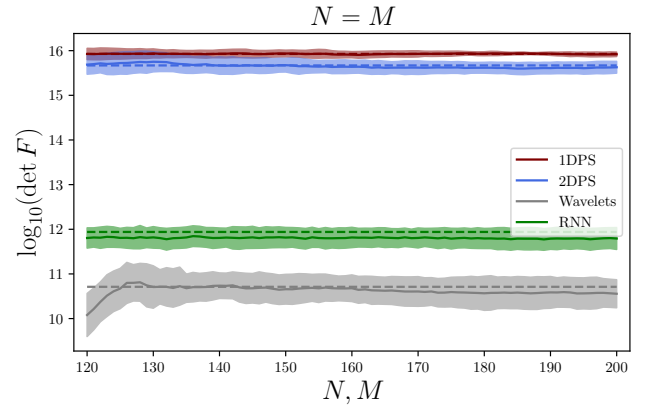

Fig. D.1. Convergence test of the Fisher information at the fiducial parameter combination, θ_{fid}, as a function of $N = M$. The solid lines (shaded regions) correspond to the means (standard deviations) computed over five random realizations of summaries at θ_{fid}. These can be compared to the dashed lines, which show the mean computed over the entire set of 1024 realizations.
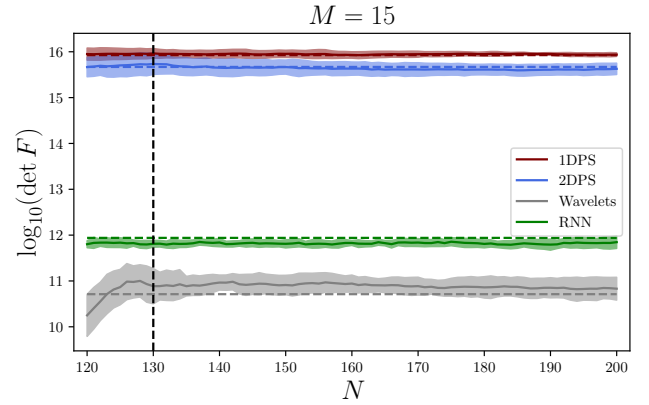

Fig. D.2. Like Fig. D.1, except here we fix $M = 15$ and vary N. The black vertical line denotes the fiducial value, $N = 130$, used in this work.
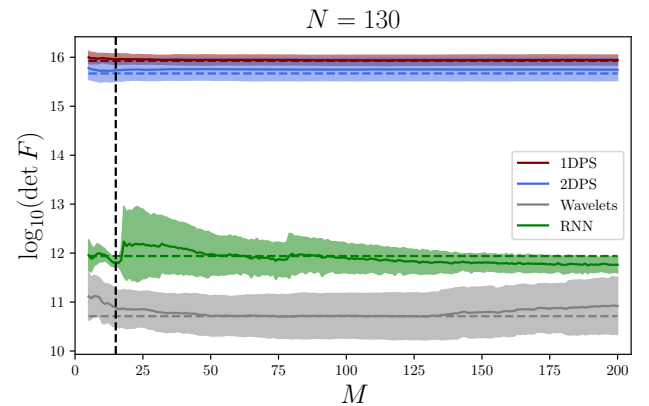

Fig. D.3. Like Fig. D.1, except here we fix $N = 130$ and vary M. The black vertical line denotes the fiducial value, $M = 15$, used in this work.