Fully Bayesian Forecasts with Evidence Networks

T. Gessey-Jones[*] and W. J. Handley[†]
Astrophysics Group, Cavendish Laboratory, J.J. Thomson Avenue, Cambridge, CB3 0HE, UK and
Kavli Institute for Cosmology, Madingley Road, Cambridge, CB3 0HA, UK

(Dated: April 23, 2024)

Sensitivity forecasts inform the design of experiments and the direction of theoretical efforts. To arrive at representative results, Bayesian forecasts should marginalize their conclusions over uncertain parameters and noise realizations rather than picking fiducial values. However, this is typically computationally infeasible with current methods for forecasts of an experiment's ability to distinguish between competing models. We thus propose a novel simulation-based methodology capable of providing expedient and rigorous Bayesian model comparison forecasts without relying on restrictive assumptions.

Keywords: Bayesian methods, Deep learning, Cosmology

Introduction.— Jeffrey and Wandelt [1] recently proposed Evidence Networks. These are classifier artificial neural networks trained on data generated from two competing models using specific loss functions, so that the output of the network converges to an invertible function of the Bayes ratio [2] between the two models. Consequently, these networks can be used for simulation-based Bayesian model comparison as illustrated in various examples by those authors. Here, we highlight a promising use case of these networks not mentioned in that original paper, forecasts of Bayesian model comparison analyses from upcoming experiments.

Forecasts are an essential part of science. Estimates of required sensitivities guide the design of experiments, and whether these experiments can probe different theoretical models informs which models time and resources are spent developing. Furthermore, funding agencies use anticipated scientific conclusions as part of their decision-making. As a result, fast, accurate, and reliable forecasting techniques play a vital role in the modern scientific method.

Due to its importance, many techniques exist to perform forecasts [e.g. 3–8]. Within a Bayesian paradigm, arguably the most accurate forecasting technique is to generate mock data and perform Bayesian analysis on the data as if it were experimental data [e.g. 9, 10]. However, such an analysis is typically computationally costly, limiting its application to a few mock data realizations. Hence, these analyses often break a central tenant of Bayesian statistics by drawing conclusions based on a small number of fiducial parameter values, which may lead to erroneous conclusions if the mock data used is not representative of reality.

To avoid this issue, the conclusions of Bayesian forecasts should be marginalized over uncertain parameters and noise realizations to perform a fully Bayesian forecast [4, 11–16]. However, performing such an analysis in practice is often infeasible with the tools used to analyse experimental data sets, e.g., MCMC and nested sampling, due to the aforementioned computational cost. Hence, Fisher forecasts [3], for parameter constraint estimates, or Savage–Dickey forecasts [4, 17], for model comparison projections, are commonly employed across astronomy and astrophysics [4, 18–29] to cover the potential data space and investigate how conclusions vary. Unfortunately, the Gaussianity assumption these forecasts rely upon does not always hold, limiting their reliability [e.g. 30, 31]. Furthermore, Savage–Dickey forecasts can only be used for nested models (when one of the models is a special case of the other), limiting their applicability [14].

Current analysis methodologies for Bayesian model comparison are thus either too slow for fully Bayesian forecasts to be feasible e.g., nested sampling, or only applicable in specific cases, e.g., Savage–Dickey forecasts. In this letter, we propose utilizing simulation-based inference in the form of Evidence Networks [1] to overcome the constraints of current methods [4, 11], enabling expedient, fully Bayesian forecasts on any scientific question formulated as a condition on the Bayes ratio between two models. Such questions include whether a signal can be detected from within noise, whether two competing theories can be distinguished by expected data, or at what sensitivity level will a piece of additional physics be required within a model? We then demonstrate the technique by finding the *a priori* chance of detection of the global 21-cm signal by a REACH-like experiment [32], an analysis that would have been computationally impracticable using traditional methods.

Bayesian Forecasting.— A traditional Bayesian analysis [see 2, for a more detailed discussion] starts with a model M that takes some parameters θ. From the model and experimental considerations, a likelihood of observing data D is constructed $\mathcal{L}(D|\theta, M)$. In addition, an initial measure, the prior $\pi(\theta|M)$, is put over the parameter space to quantify the *a priori* knowledge of the parameter values. Then given some (mock) observed data, the measure over the parameter space is updated via an

[*] tg400@cam.ac.uk
[†] wh260@mrao.cam.ac.uk

application of Bayes' theorem

$$\mathcal{P}(\theta|D,M) = \frac{\mathcal{L}(D|\theta,M)\pi(\theta|M)}{\mathcal{Z}}, \tag{1}$$

into the *a posteriori* knowledge of the parameter values $\mathcal{P}(\theta|D,M)$, called the posterior. Here $\mathcal{Z}$ is the Bayesian evidence

$$\mathcal{Z} = \int \mathcal{L}(D|\theta,M)\pi(\theta|M)d\theta = P(D|M), \tag{2}$$

which serves both as a normalization constant and as a natural goodness-of-fit statistic for comparing competing models of the same data. Now suppose there are competing models, M_0 and M_1, with corresponding Bayesian evidences $\mathcal{Z}_\mathrm{i}$. If we have an *a priori* belief in each model of $P(M_\mathrm{i})$ then for each model the *a posteriori* belief $P(M_\mathrm{i}|D)$ in the model is found using a further application of Bayes' theorem

$$P(M_\mathrm{i}|D) = \frac{P(D|M_\mathrm{i})P(M_\mathrm{i})}{P(D)} = \frac{\mathcal{Z}_\mathrm{i}P(M_\mathrm{i})}{P(D)}. \tag{3}$$

Cancelling $P(D)$ between the above with $i = 0$ and $i = 1$ then gives the Bayesian Model Comparison equation

$$\frac{P(M_1|D)}{P(M_0|D)} = \frac{\mathcal{Z}_1 P(M_1)}{\mathcal{Z}_0 P(M_0)}. \tag{4}$$

This equation shows that after observing some data D the relative belief in the two models is updated with the ratio of their evidences, e.g. the belief in the higher evidence model increases, and the belief in the lower evidence model belief decreases. It is common to assume initially that the two models are equally likely $P(M_1) = P(M_0)$ in which case equation (4) simplifies to

$$\mathcal{K} = \frac{P(M_1|D)}{P(M_0|D)} = \frac{\mathcal{Z}_1}{\mathcal{Z}_0}, \tag{5}$$

so that the Bayes ratio $\mathcal{K}$ of the model evidences is simply the relative belief in model 1 compared to model 0 after observing some data. The above can also be rearranged to find the posterior belief in a model

$$P(M_1|D) = \frac{\mathcal{K}}{1+\mathcal{K}}, \tag{6}$$

which can be compared to a statistical threshold to determine if one model is significantly preferred over the other.

With the mathematical notation of Bayesian analysis established, we can return to the central topic of this letter; forecasts. For many scientific applications [11, 14], the question of interest is best posed as whether a proposed experiment could distinguish between two competing models. A traditional Bayesian forecast of such question would go through the following steps. First, the models and corresponding likelihoods are established. Then, mock data is simulated from one of the models by choosing some fiducial parameter values $\tilde{\theta}$, adding noise η, and experimental effects. $\mathcal{Z}$ is evaluated for each model treating the mock data as you would actual observations from which $\mathcal{K}$ (or equivalently $P(M_\mathrm{i}|D)$) is calculated. Finally, $\mathcal{K}$ is compared to some condition to determine if the favouring of a model is statistically significant and so the models are distinguishable.

The conclusion of the outlined forecast is hence based on a condition on $\mathcal{K}$ of the form $\mathcal{K} > \mathcal{K}_\mathrm{crit}$ with $\mathcal{K}_\mathrm{crit}$ a statistical significance threshold. However, $\mathcal{K}$ is conditional on the fiducial parameters chosen to generate the mock data and the noise added, so it is more accurate to state

$$\mathcal{K}\left(\tilde{\theta},\eta\right) > \mathcal{K}_\mathrm{crit}. \tag{7}$$

Using such a criterion presents two major issues (previously identified and discussed in Mukherjee *et al.* [11] and Trotta [4]). Firstly the result is dependent on random noise η, leading to differences in conclusion purely based on chance. Secondly, these conclusions stem from a single point in the parameter space $\tilde{\theta}$, leading to differences based on the choice of fiducial model (anathema to Bayesian statistics). A central tenant of Bayesian statistics being that particular sets of parameter values, even those that maximize the likelihood or posterior, have zero measure and thus are vanishingly unlikely and should not be used to draw conclusions. Instead, in Bayesian statistics, conclusions should be reached by marginalizing over the parameter space weighted by some measure, typically the prior or posterior. This issue is further compounded in the case of forecasts as there is often limited advance knowledge of the parameter values leading to the conclusion drawn potentially varying widely between choices of $\tilde{\theta}$.

A rigorously Bayesian approach would be to marginalize the condition in equation (7) over $\tilde{\theta}$ and η. Such a procedure would give the expected probability of the condition holding and thus of drawing some scientific conclusion rather than a binary yes or no answer. Since the prior represents our knowledge of the parameters before any data is measured, it is the natural measure for a forecasting analysis. So, for our example, we should calculate

$$\mathbb{E}(\text{Distinguish Models}) = \left\langle \mathcal{K}\left(\tilde{\theta},\eta\right) > \mathcal{K}_\mathrm{crit} \right\rangle_{\eta,\pi(\tilde{\theta})}, \tag{8}$$

to find a fully Bayesian forecast of our expected chances of distinguishing between the models at the statistical significance set by $\mathcal{K}_\mathrm{crit}$. More generally, the condition in equation (8) could be replaced with any other condition on $\mathcal{K}$ to draw a scientific conclusion

$$\mathbb{E}(\text{Drawing Conclusion}) = \left\langle \text{Condition}\left[\mathcal{K}(\tilde{\theta},\eta)\right] \right\rangle_{\eta,\pi(\tilde{\theta})}, \tag{9}$$

to perform a fully Bayesian forecast on drawing said conclusion. In addition, equation (9) can be modified to produce fully Bayesian forecasts for a broader range of scientific questions. For example, when appropriate to the question being posed, the marginalization should be performed over the model prior as well (this is equivalent to marginalizing over the predictive posterior odds distribution introduced in Trotta [4]), or over conditional priors. Thus, fully Bayesian forecasts can in theory be used to give principled and interpretable answers to a range of forecasting questions.

However, performing the above average over the parameters and noise (and potentially models) in practice would require thousands or millions of mock data sets and, in turn, thousands or millions of Bayes ratio calculations. Mukherjee *et al.* [11] proposed using Nested Sampling [33, 34] to calculate the $\mathcal{K}$ values for fully Bayesian model comparison forecasts, but in practice, this is typically prohibitively computationally expensive, even for simple problems. To circumvent these computational limitations Trotta [14] proposed using the Savage–Dickey density ratio [17, 35] to rapidly evaluate $\mathcal{K}$ values between nested models, facilitating a fully Bayesian forecast of whether the *Planck* satellite could detect a deviation in the scalar spectral index of primordial perturbations n_s from 1. The use of the Savage–Dickey density ratio for model comparison forecasts is analogous to the widespread usage of Fisher forecasts to perform parameter constraint forecasts over uncertain data spaces. Both techniques utilize the analytic results available for linear models and Gaussian likelihoods to calculate approximate analytic parameter posteriors [e.g. 18] or the Bayes ratio between models. However, the reliability of these results is conditional on the accuracy of the implicit linearization and Gaussianity assumptions [see 30, 31]. Furthermore, usage of the Savage–Dickey density ratio requires the models being compared to be nested. Thus, if we had nested models, an explicit likelihood, and knew the above assumptions to hold well the Savage–Dickey density ratio would suffice to perform fully Bayesian model comparison forecasts. However, these requirements significantly limit the usage of such a methodology. Reliable and widely applicable fully Bayesian forecasts will hence require a novel methodology that maintains the expedience of Savage–Dickey forecasts but is applicable to non-nested models and avoids the same assumptions.

Evidence Networks.— Evidence networks are a type of classifier artificial neural networks introduced recently in Jeffrey and Wandelt [1]. They take in simulated data D and output a single value $f_\mathrm{EN}(D)$, with training performed on data generated from two models (labels $m = 1$ and $m = 0$). The principal insight with these networks is that for a broad category of choices of loss function the evidence network's value converges toward an invertible function of the Bayes ratio $\mathcal{K}(D)$ between model 1 and model 0. Hence, if such a network is converged, the Bayes ratio between the two models for any set of 'observed' data can be directly calculated from the network's output.

As established in the previous section, efficient evaluation of $\mathcal{K}$ for a wide range of mock data is the main obstacle to performing practicable fully Bayesian forecasts. Since evaluating a neural network on a GPU is orders of magnitude faster than direct Bayesian evidence calculation techniques, utilizing evidence networks may circumvent this computational limitation. We thus propose an alternative evidence-network-based methodology for performing fully Bayesian forecasts:

- First, create simulators of mock data from the two competing models (including experimental considerations such as noise or selection effects).
- Then, using the two simulators, generate a training set and validation set of labelled mock data and train the evidence network.
- Validate the evidence network using a blind coverage test [1] to verify the network's accuracy and, if possible, also compare its output to a sample of $\mathcal{K}$ values calculated from traditional Bayesian techniques. This step is essential, as for data manifolds too complex to be classified by the chosen network architecture, or for insufficient training data sets, the converged network will not correspond to an accurate calculation of $\mathcal{K}$[1]. If validation indicates the network has not accurately converged, the network architecture or training will need to be refined and the previous steps repeated.
- Finally, evaluate equation (9), or equivalent, using the network and previously developed simulators to evaluate $\mathcal{K}$ over π and η (and potentially M_i) efficiently.

We anticipate this approach to be highly performant relative to the traditional Bayesian approach since each $\mathcal{K}$ evaluation no longer requires an exploration of the parameter space. Furthermore, our methodology does not require the models used to be nested or to satisfy the restrictive assumptions of the Savage–Dickey forecasting method; and only needs efficient mock data simulators rather than explicit likelihoods. As a result, it can even be applied in simulation-based inference contexts where no closed-form likelihood exists. We thus also anticipate it to be widely applicable.

An Application to Cosmology.— We have motivated fully Bayesian forecasts and argued that evidence networks may make performing one feasible. Here, we demonstrate this feasibility while also illustrating some of the insights gained from such an analysis.

The problem we tackle is determining the expected chances of a 21-cm global signal experiment making

[1] In the example discussed in the next section we found a blind coverage test revealed the network was consistently underconfident (e.g., conservative) in these two scenarios.

a definitive detection. Through the redshift evolution of the 21-cm global signal the thermal evolution of the universe from recombination to reionization can be traced, giving insights into cosmology and structure formation [see 36, for a more detailed introduction to the field]. However, measurement of the signal is challenging as its expected magnitude is five orders of magnitude below that of galactic foregrounds. As a result, there are currently no definitive detections of the global 21-cm signal. The EDGES 2 experiment claimed a signal detection [37], but this is disputed [e.g. 38] due to the signal not matching theoretical expectations [e.g. 39], being better fit by the presence of a systematic [e.g. 40], and the null-detection from the SARAS 3 experiment [41]. Because of this lack of definitive detection, there are several ongoing and proposed experiments to try and measure the sky-averaged 21-cm signal, e.g. REACH [32], PRIZM [42], and EDGES 3. The question then naturally arises, what is the *a priori* expectation of a global signal experiment with given sensitivity making a definitive detection of the uncertain 21-cm signal? Since a significant signal detection can be determined to occur when $\mathcal{K}$ between a model with a signal and a model without a signal exceeds some threshold $\mathcal{K}_{\mathrm{crit}}$, this question can be rigorously answered through a fully Bayesian forecast of the form given in equation (8).

Hence, following our fully Bayesian forecasting methodology outlined above, we began by constructing the two mock data simulators, one that modelled a global 21-cm signal and one without. We imagined a REACH-like global 21-cm signal experiment with frequency-band covering redshift 7.5 to 28.0 and an analysis spectral resolution of 1 MHz. For both simulators, we used the physically motivated galactic and ionospheric foreground model from Hills *et al.* [38], and assumed Gaussian white noise with magnitude $\sigma_{\mathrm{noise}} = 0.015\,\mathrm{K}$[2]. For our global 21-cm signal model, we used GLOBALEMU [43], an emulator of a more computationally costly semi-numerical simulation code [e.g., 44–46]. This model has seven parameters, star formation efficiency f_*, minimum circular virial velocity V_{c}, X-ray emission efficiency f_{X}, optical depth to the cosmic microwave background τ, exponent α and lower-cutoff E_{min} of the X-ray spectral energy distribution, and the maximum root-mean free path of ionizing photons R_{mfp}. We specified our *a priori* knowledge of these parameters and the foreground parameters through our priors listed in table I. The astrophysical priors used are uniform or log-uniform distributions centred on theoretically expected values, except for τ, which we used a truncated Gaussian prior based on the *Planck* 2018 posterior on τ [47]. For the foreground parameter priors, we used physically restricted priors following Hills *et al.* [38]. Thus by combining noise generators, our analytic foreground model, GLOBALEMU, and samplers over our priors, we constructed simulators of mock global 21-cm signal data from a model with only noise and foreground (no-signal) and a model with noise, foreground, and signal (with-signal). Our two competing models thus have moderate dimensionalities (5 and 12) and depend non-linearly on their parameters, leading to complex data manifolds. In addition, this particular classification task is made more challenging due to the large shared foreground component of the two models being 10^4 to 10^5 times larger than the signal.

TABLE I. Priors on our foreground and 21-cm global signal parameters. To keep the parameter values within the training parameter ranges of GLOBALEMU we use a truncated Gaussian prior on τ derived from the Planck 2018 measurements rather than a Gaussian prior.

Parameter	Prior Type	Min	Max	Mean	Std.
d_0 (K)	Uniform	1500	2000	-	-
d_1	Uniform	-1.0	1.0	-	-
d_2	Uniform	-0.05	0.05	-	-
τ_e	Uniform	0.005	0.200	-	-
T_e (K)	Uniform	200	2000	-	-
f_*	Log Uniform	0.0001	0.5	-	-
V_{c} ($\mathrm{km\,s^{-1}}$)	Log Uniform	4.2	30.0	-	-
f_{x}	Log Uniform	0.001	1000.0	-	-
τ	Truncated Gaussian	0.040	0.17	0.054	0.007
α	Uniform	1.0	1.5	-	-
E_{min} (keV)	Log Uniform	0.1	3.0	-	-
R_{mfp} (cMpc)	Uniform	10.0	50.0	-	-

To train our evidence network, we then generated 32,000,000 (12,800,000) mock data sets from each simulator to form our training (validation) set. Our evidence network was implemented in TENSORFLOW [48], and consisted of an initial Cholesky whitening transform [49] followed by dense hidden layers of size 256-256-64-64-64-64-64-64-1, with batch normalization and a ReLU activation functions on all layers except for the output node, and an additive skip connection between the third and sixth layers to ease training. We used the $\alpha = 2$ l-POP exponential loss function recommended for evident networks by Jeffrey and Wandelt [1] and the Adam optimizer with an initial learning rate of 10^{-3}, decay steps of 10^5, and decay rate of 0.95. The network was trained with early-stopping for 900 epochs using a batch size of 32,768.

To validate the network had converged to an accurate prediction of $\mathcal{K}$, we performed a blind coverage test as outlined in Jeffrey and Wandelt [1]. This test consists of using the network to predict the posterior model probabilities for a range of test data sets, then binning the data sets by these probabilities. If the network has correctly converged, the model probability corresponding to each bin should equal the proportion of the data sets in the bin generated from said model, which we indeed find to be the case for our testing set composed of 1,000,000 mock data sets generated from each model. Furthermore,

[2] As the noise scales with spectral resolution $\Delta\nu$ as $1/\sqrt{\Delta\nu}$ this is equivalent to 0.047 K noise at a resolution of 0.1 MHz, which lies between the *pessimistic* and *expected* case projected for REACH [32].

as in this case, we can construct an explicit likelihood, we additionally validated our network and methodology by comparing whether a significant detection would be concluded based on the network $\mathcal{K}$ values and $\mathcal{K}$ computed using POLYCHORD [50, 51]. For a signal detection threshold of $3\,\sigma$ ($5\,\sigma$)[3], corresponding to $\log(\mathcal{K}_{\mathrm{crit}}) = 5.91$ ($\log(\mathcal{K}_{\mathrm{crit}}) = 14.4$), we found the two methods came to the same conclusion on whether a signal was detected for 96.6% (95.1%) of 1000 mock data sets from our noisy-signal model, showing good agreement.

Finally, we evaluated equation (8) using our evidence network and 1,000,000 sets of noisy-signal model mock data. Ultimately, we found the expected chance of the experiment detecting the global 21-cm signal at a statistical significance of $3\,\sigma$ ($5\,\sigma$) was 46.0% (32.4%). This approximately 50% change of a 21-cm signal detection at $> 3\,\sigma$ confidence, suggests the 0.015 K sensitivity at 1 MHz resolution considered here is indicative of the minimum sensitivity global signal experiments such target.

Additionally, our 1,000,000 $\mathcal{K}$ evaluations allow for a broad range of further analyses at little to no additional computation cost. Instead of performing the average over π in equation (8), we can marginalize over conditional priors with one or two parameters fixed, giving insight into under which early Universe astrophysical scenarios we would expect to detect the 21-cm signal. These conditional detection probabilities are depicted in figure 1 for the parameters with which strong variation in detection probability was seen, f_*, f_{X}, and τ. We find the detection of the 21-cm signal is more likely for high star formation efficiencies, X-ray efficiencies around the theoretically expected value of 1, and higher optical depths to reionization. With an almost 100% chance of a $3\,\sigma$ detection for $f_* > 0.01$ and $f_{\mathrm{X}} = 1$, and an effectively 0% chance of a $3\,\sigma$ detection for $f_* < 0.001$ or $f_{\mathrm{X}} > 30$. This strong variation in definitive detection chances retroactively provides further motivation for fully Bayesian forecasts as we see the conclusion drawn would be highly sensitive to the fiducial global 21-cm signal parameters chosen for a traditional Bayesian forecast.

There are a myriad of further analyses we could perform with our methodology. For example, we could study the evolution of the detection probability against the experiment noise level or bandwidth, which could, in turn, help inform experimental design. Additionally, we could consider the functional distribution of the detected and non-detected 21-cm global signals to gain insight into what differentiates these two categories. Or alternatively, we could investigate how the chances of a 21-cm global signal detection have changed in light of the parameter constraints from the HERA 21-cm power spectrum limits [52, 53], or the SARAS 3 null detection [41, 54].

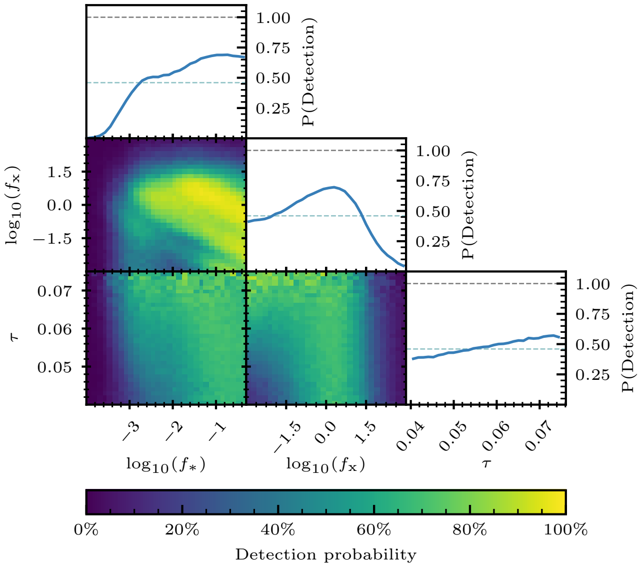

FIG. 1. Triangle plot depicting the probability of a global 21-cm signal detection at $\geq 3\,\sigma$ statistical significance, by an experiment covering redshift 7.5 to 28.0, with frequency resolution $\Delta\nu = 1\,\mathrm{MHz}$, assuming the Hills *et al.* [38] physical foreground model and white noise of $\sigma_{\mathrm{noise}} = 0.015\,\mathrm{K}$. The diagonal shows the total probability of a detection marginalized over noise realizations and the parameter space except for one fixed parameter, and the below diagonal is the equivalent with two fixed parameters. The total detection probability with no parameters fixed was 46.0%. V_{c}, α, E_{min}, and R_{mfp} are not shown because the probability of detection was found to vary only weakly with them. We find high f_*, $f_{\mathrm{X}} \approx 1$, and high τ values increase the chance of a 21-cm signal detection, with f_* the most impactful. For different parameter combinations, we find detection probabilities varying from $\sim 0\%$ to $\sim 100\%$, illustrating the need for fully Bayesian forecasts as the conclusion of detectability varies strongly over the *a priori* uncertain parameter space.

Furthermore, if this were a problem where particular parameter(s) values were of special interest (e.g., the minimum sum of the neutrino masses permitted by particle physics) this methodology also allows for determining detection chance with the parameter(s) fixed to that value while still marginalizing over noise and nuisance parameters. However, we shall leave such analyses to future work since the focus of this letter is on the method and its feasibility rather than particular scientific problems.

Let us return now to the computational performance of our method. The training data generation, network training, forecast data generation, network $\mathcal{K}$ evaluations, and plotting of figure 1, took a combined total of 5.54 GPU hours[4]. Conversely, the 1000 POLYCHORD evaluations of $\mathcal{K}$ as part of our validation process took a total

[3] These σ thresholds are translated to $\mathcal{K}_{\mathrm{crit}}$ values via equation (6), and requiring that the posterior probability of the model with no signal is less than the corresponding p values of 2.70×10^{-3} or 5.73×10^{-7}.

[4] On an NVIDIA A100-SXM-80GB GPU that was part of a CSD3 HPC Ampere GPU node.

of 45,000 CPU hours[5], from which we can estimate the 1,000,000 $\mathcal{K}$ evaluations used in our fully Bayesian forecast would have required 45,000,000 CPU hours using traditional methods. While it is not meaningful to directly compare GPU to CPU hours we can compare the costs of those hours. On the cluster we utilized GPU hours are charged at 50 times the rate of CPU hours. Thus our method gives a cost-weighted performance improvement of $10^{5.2}$. Since this performance gain was for a single problem, and our implementation was not optimized, this level of performance gain cannot be assumed to apply universally. However, it is indicative our methodology is highly performant compared to traditional techniques and, as we have directly demonstrated, facilitates analyses that were previously computationally prohibitively expensive.

Conclusions.— We have argued, like Mukherjee *et al.* [11] and Trotta [4] before us, that to arrive at accurate and interpretable predictions, the conclusions of scientific forecasts should be marginalized over any uncertain model parameters and noise realizations. However, such fully Bayesian forecasts are computationally infeasible with traditional methods for model comparison forecasts. We thus propose a novel methodology for performing fully Bayesian forecasts based on Evidence Networks.

To illustrate our method and the insights that can be gained from fully Bayesian forecasts, we applied it to determine the chances of a REACH-like experiment detecting the global 21-cm signal from beneath foregrounds and noise. For a frequency resolution of 1 MHz and a noise level of $\sigma_{\text{noise}} = 0.015\,\text{K}$, we find a 46.0% (32.4%) chance of detection at $3\,\sigma$ ($5\,\sigma$). Thus suggesting this noise level is indicative of the minimum sensitivity global 21-cm signal experiments should target. Additionally, our methodology allows us to produce triangle plots of how this chance of detection varies when one or two model parameters are fixed, at no extra computational cost. For this example problem, we find a cost-weighted speed-up of $10^{5.2}$ using our approach compared to a traditional nested-sampling-based method that would have taken $45,000,000$ CPU hours.

The method we propose can be applied to any forecasting question which can be formulated as a condition on the Bayes ratio between two models. This includes: if a signal can be detected from within noise (e.g. gravitational waves [55], or the 21-cm signal [36]); whether two competing theories can be distinguished by anticipated data (e.g. MOND [56] or General Relativity [57]); or if the inclusion of novel physics in a model will be necessary (e.g. neutrinos in CMB experiment analysis [26]). Additionally, the method only requires simulators of mock data, and thus can still be used in cases where closed-form likelihoods or explicit priors are not available. As a result, this methodology should allow for reliable and efficient fully Bayesian forecasts on a wide range of forecasting problems.

Since the proposed methodology is simulation-based and has a low computational cost, we anticipate it will be particularly suited to performing forecasts for a range of potential experimental configurations. Thus allowing for the optimization of experimental configurations to minimize cost or maximize the chance of the detection of new physics. To facilitate the application of this methodology by others, we make public on GitHub all codes and data used in the writing of this letter.

ACKNOWLEDGMENTS

We would like to express our gratitude to Harry Bevins for several insightful conversations concerning this work. The authors would like to thank the Science and Technology Facilities Council (UK) for their support of TGJ through grant number ST/V506606/1, and the Royal Society for their support of WJH through a Royal Society University Research Fellowship. This work was performed using the Cambridge Service for Data Driven Discovery (CSD3), part of which is operated by the University of Cambridge Research Computing on behalf of the STFC DiRAC HPC Facility (www.dirac.ac.uk). The DiRAC component of CSD3 was funded by BEIS capital funding via STFC capital grants ST/P002307/1 and ST/R002452/1 and STFC operations grant ST/R00689X/1. DiRAC is part of the National e-Infrastructure. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

[1] N. Jeffrey and B. D. Wandelt, Machine Learning: Science and Technology **5**, 015008 (2024), arXiv:2305.11241 [cs.LG].

[2] D. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, 2003).

[3] R. A. Fisher, Philosophical Transactions of the Royal Society of London Series A **222**, 309 (1922).

[4] R. Trotta, MNRAS **378**, 819 (2007), arXiv:astro-ph/0703063 [astro-ph].

[5] E. Sellentin, M. Quartin, and L. Amendola, MNRAS **441**, 1831 (2014), arXiv:1401.6892 [astro-ph.CO].

[6] J. Alvey, M. Escudero, and N. Sabti, J. Cosmology Astropart. Phys. **2022**, 037 (2022), arXiv:2111.12726 [astro-ph.CO].

[7] J. Alvey, M. Escudero, N. Sabti, and T. Schwetz,

[5] On 76 Intel Ice Lake CPUs that form a CSD3 HPC Ice Lake node.

Phys. Rev. D **105**, 063501 (2022), arXiv:2111.14870 [hep-ph].

[8] J. Ryan, B. Stevenson, C. Trendafilova, and J. Meyers, Phys. Rev. D **107**, 103506 (2023), arXiv:2211.06534 [astro-ph.CO].

[9] D. Anstey, E. de Lera Acedo, and W. Handley, MNRAS **506**, 2041 (2021), arXiv:2010.09644 [astro-ph.IM].

[10] S. Rieck, A. W. Criswell, V. Korol, M. A. Keim, M. Bloom, and V. Mandic, arXiv e-prints , arXiv:2308.12437 (2023), arXiv:2308.12437 [astro-ph.IM].

[11] P. Mukherjee, D. Parkinson, P. S. Corasaniti, A. R. Liddle, and M. Kunz, MNRAS **369**, 1725 (2006), arXiv:astro-ph/0512484 [astro-ph].

[12] D. S. Sivia and J. Skilling, *Data analysis: a Bayesian tutorial*, 2nd ed., Oxford science publications (Oxford University Press, Oxford, 2006).

[13] C. Pahud, A. R. Liddle, P. Mukherjee, and D. Parkinson, Phys. Rev. D **73**, 123524 (2006), arXiv:astro-ph/0605004 [astro-ph].

[14] R. Trotta, MNRAS **378**, 72 (2007), arXiv:astro-ph/0504022 [astro-ph].

[15] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis* (Chapman and Hall/CRC, 2014).

[16] J. S. Y. Leung and Z. Huang, International Journal of Modern Physics D **26**, 1750070 (2017), arXiv:1604.07330 [astro-ph.CO].

[17] J. M. Dickey, The Annals of Mathematical Statistics **42**, 204 (1971).

[18] M. Tegmark, A. N. Taylor, and A. F. Heavens, ApJ **480**, 22 (1997), arXiv:astro-ph/9603021 [astro-ph].

[19] A. Albrecht, G. Bernstein, R. Cahn, W. L. Freedman, J. Hewitt, W. Hu, J. Huth, M. Kamionkowski, E. W. Kolb, L. Knox, J. C. Mather, S. Staggs, and N. B. Suntzeff, arXiv e-prints , astro-ph/0609591 (2006), arXiv:astro-ph/0609591 [astro-ph].

[20] H.-J. Seo and D. J. Eisenstein, ApJ **665**, 14 (2007), arXiv:astro-ph/0701079 [astro-ph].

[21] M. Vallisneri, Phys. Rev. D **77**, 042001 (2008), arXiv:gr-qc/0703086 [gr-qc].

[22] S. More, F. C. van den Bosch, M. Cacciato, A. More, H. Mo, and X. Yang, MNRAS **430**, 747 (2013), arXiv:1207.0004 [astro-ph.CO].

[23] E. Di Dio, F. Montanari, R. Durrer, and J. Lesgourgues, J. Cosmology Astropart. Phys. **2014**, 042 (2014), arXiv:1308.6186 [astro-ph.CO].

[24] Z. Zhai and M. R. Blanton, ApJ **850**, 41 (2017), arXiv:1707.06555 [astro-ph.CO].

[25] C. Bonvin and P. Fleury, J. Cosmology Astropart. Phys. **2018**, 061 (2018), arXiv:1803.02771 [astro-ph.CO].

[26] P. Ade, J. Aguirre, Z. Ahmed, S. Aiola, A. Ali, D. Alonso, M. A. Alvarez, K. Arnold, P. Ashton, J. Austermann, H. Awan, C. Baccigalupi, T. Baildon, D. Barron, N. Battaglia, R. Battye, E. Baxter, A. Bazarko, J. A. Beall, R. Bean, D. Beck, S. Beckman, B. Beringue, F. Bianchini, S. Boada, D. Boettger, J. R. Bond, J. Borrill, M. L. Brown, S. M. Bruno, S. Bryan, E. Calabrese, V. Calafut, P. Calisse, J. Carron, A. Challinor, G. Chesmore, Y. Chinone, J. Chluba, H.-M. S. Cho, S. Choi, G. Coppi, N. F. Cothard, K. Coughlin, D. Crichton, K. D. Crowley, K. T. Crowley, A. Cukierman, J. M. D'Ewart, R. Dünner, T. de Haan, M. Devlin, S. Dicker, J. Didier, M. Dobbs, B. Dober, C. J. Duell, S. Duff, A. Duivenvoorden, J. Dunkley, J. Dusatko, J. Errard, G. Fabbian, S. Feeney, S. Ferraro, P. Fluxà, K. Freese, J. C. Frisch, A. Frolov, G. Fuller, B. Fuzia, N. Galitzki, P. A. Gallardo, J. Tomas Galvez Ghersi, J. Gao, E. Gawiser, M. Gerbino, V. Gluscevic, N. Goeckner-Wald, J. Golec, S. Gordon, M. Gralla, D. Green, A. Grigorian, J. Groh, C. Groppi, Y. Guan, J. E. Gudmundsson, D. Han, P. Hargrave, M. Hasegawa, M. Hasselfield, M. Hattori, V. Haynes, M. Hazumi, Y. He, E. Healy, S. W. Henderson, C. Hervias-Caimapo, C. A. Hill, J. C. Hill, G. Hilton, M. Hilton, A. D. Hincks, G. Hinshaw, R. Hložek, S. Ho, S.-P. P. Ho, L. Howe, Z. Huang, J. Hubmayr, K. Huffenberger, J. P. Hughes, A. Ijjas, M. Ikape, K. Irwin, A. H. Jaffe, B. Jain, O. Jeong, D. Kaneko, E. D. Karpel, N. Katayama, B. Keating, S. S. Kernasovskiy, R. Keskitalo, T. Kisner, K. Kiuchi, J. Klein, K. Knowles, B. Koopman, A. Kosowsky, N. Krachmalnicoff, S. E. Kuenstner, C.-L. Kuo, A. Kusaka, J. Lashner, A. Lee, E. Lee, D. Leon, J. S. Y. Leung, A. Lewis, Y. Li, Z. Li, M. Limon, E. Linder, C. Lopez-Caraballo, T. Louis, L. Lowry, M. Lungu, M. Madhavacheril, D. Mak, F. Maldonado, H. Mani, B. Mates, F. Matsuda, L. Maurin, P. Mauskopf, A. May, N. McCallum, C. McKenney, J. McMahon, P. D. Meerburg, J. Meyers, A. Miller, M. Mirmelstein, K. Moodley, M. Munchmeyer, C. Munson, S. Naess, F. Nati, M. Navaroli, L. Newburgh, H. N. Nguyen, M. Niemack, H. Nishino, J. Orlowski-Scherer, L. Page, B. Partridge, J. Peloton, F. Perrotta, L. Piccirillo, G. Pisano, D. Poletti, R. Puddu, G. Puglisi, C. Raum, C. L. Reichardt, M. Remazeilles, Y. Rephaeli, D. Riechers, F. Rojas, A. Roy, S. Sadeh, Y. Sakurai, M. Salatino, M. Sathyanarayana Rao, E. Schaan, M. Schmittfull, N. Sehgal, J. Seibert, U. Seljak, B. Sherwin, M. Shimon, C. Sierra, J. Sievers, P. Sikhosana, M. Silva-Feaver, S. M. Simon, A. Sinclair, P. Siritanasak, K. Smith, S. R. Smith, D. Spergel, S. T. Staggs, G. Stein, J. R. Stevens, R. Stompor, A. Suzuki, O. Tajima, S. Takakura, G. Teply, D. B. Thomas, B. Thorne, R. Thornton, H. Trac, C. Tsai, C. Tucker, J. Ullom, S. Vagnozzi, A. van Engelen, J. Van Lanen, D. D. Van Winkle, E. M. Vavagiakis, C. Vergès, M. Vissers, K. Wagoner, S. Walker, J. Ward, B. Westbrook, N. Whitehorn, J. Williams, J. Williams, E. J. Wollack, Z. Xu, B. Yu, C. Yu, F. Zago, H. Zhang, N. Zhu, and Simons Observatory Collaboration, J. Cosmology Astropart. Phys. **2019**, 056 (2019), arXiv:1808.07445 [astro-ph.CO].

[27] Euclid Collaboration, A. Blanchard, S. Camera, C. Carbone, V. F. Cardone, S. Casas, S. Clesse, S. Ilić, M. Kilbinger, T. Kitching, M. Kunz, F. Lacasa, E. Linder, E. Majerotto, K. Markovič, M. Martinelli, V. Pettorino, A. Pourtsidou, Z. Sakr, A. G. Sánchez, D. Sapone, I. Tutusaus, S. Yahia-Cherif, V. Yankelevich, S. Andreon, H. Aussel, A. Balaguera-Antolínez, M. Baldi, S. Bardelli, R. Bender, A. Biviano, D. Bonino, A. Boucaud, E. Bozzo, E. Branchini, S. Brau-Nogue, M. Brescia, J. Brinchmann, C. Burigana, R. Cabanac, V. Capobianco, A. Cappi, J. Carretero, C. S. Carvalho, R. Casas, F. J. Castander, M. Castellano, S. Cavuoti, A. Cimatti, R. Cledassou, C. Colodro-Conde, G. Congedo, C. J. Conselice, L. Conversi, Y. Copin, L. Corcione, J. Coupon, H. M. Courtois, M. Cropper, A. Da Silva, S. de la Torre, D. Di Ferdinando, F. Dubath, F. Ducret, C. A. J. Duncan, X. Dupac, S. Dusini, G. Fabbian, M. Fabricius, S. Farrens, P. Fosalba, S. Fotopoulou, N. Fourmanoit, M. Frailis,

E. Franceschi, P. Franzetti, M. Fumana, S. Galeotta, W. Gillard, B. Gillis, C. Giocoli, P. Gómez-Alvarez, J. Graciá-Carpio, F. Grupp, L. Guzzo, H. Hoekstra, F. Hormuth, H. Israel, K. Jahnke, E. Keihanen, S. Kermiche, C. C. Kirkpatrick, R. Kohley, B. Kubik, H. Kurki-Suonio, S. Ligori, P. B. Lilje, I. Lloro, D. Maino, E. Maiorano, O. Marggraf, N. Martinet, F. Marulli, R. Massey, E. Medinaceli, S. Mei, Y. Mellier, B. Metcalf, J. J. Metge, G. Meylan, M. Moresco, L. Moscardini, E. Munari, R. C. Nichol, S. Niemi, A. A. Nucita, C. Padilla, S. Paltani, F. Pasian, W. J. Percival, S. Pires, G. Polenta, M. Poncet, L. Pozzetti, G. D. Racca, F. Raison, A. Renzi, J. Rhodes, E. Romelli, M. Roncarelli, E. Rossetti, R. Saglia, P. Schneider, V. Scottez, A. Secroun, G. Sirri, L. Stanco, J. L. Starck, F. Sureau, P. Tallada-Crespí, D. Tavagnacco, A. N. Taylor, M. Tenti, I. Tereno, R. Toledo-Moreo, F. Torradeflot, L. Valenziano, T. Vassallo, G. A. Verdoes Kleijn, M. Viel, Y. Wang, A. Zacchei, J. Zoubian, and E. Zucca, A&A **642**, A191 (2020), arXiv:1910.09273 [astro-ph.CO].

[28] W. d'Assignies D, C. Zhao, J. Yu, and J.-P. Kneib, MNRAS **521**, 3648 (2023), arXiv:2301.02289 [astro-ph.CO].

[29] C. A. Mason, J. B. Muñoz, B. Greig, A. Mesinger, and J. Park, MNRAS **524**, 4711 (2023), arXiv:2212.09797 [astro-ph.CO].

[30] L. Perotto, J. Lesgourgues, S. Hannestad, H. Tu, and Y. Y Y Wong, J. Cosmology Astropart. Phys. **2006**, 013 (2006), arXiv:astro-ph/0606227 [astro-ph].

[31] L. Wolz, M. Kilbinger, J. Weller, and T. Giannantonio, J. Cosmology Astropart. Phys. **2012**, 009 (2012), arXiv:1205.3984 [astro-ph.CO].

[32] E. de Lera Acedo, D. I. L. de Villiers, N. Razavi-Ghods, W. Handley, A. Fialkov, A. Magro, D. Anstey, H. T. J. Bevins, R. Chiello, J. Cumner, A. T. Josaitis, I. L. V. Roque, P. H. Sims, K. H. Scheutwinkel, P. Alexander, G. Bernardi, S. Carey, J. Cavillot, W. Croukamp, J. A. Ely, T. Gessey-Jones, Q. Gueuning, R. Hills, G. Kulkarni, R. Maiolino, P. D. Meerburg, S. Mittal, J. R. Pritchard, E. Puchwein, A. Saxena, E. Shen, O. Smirnov, M. Spinelli, and K. Zarb-Adami, Nature Astronomy **6**, 984 (2022).

[33] J. Skilling, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, American Institute of Physics Conference Series, Vol. 735, edited by R. Fischer, R. Preuss, and U. V. Toussaint (2004) pp. 395–405.

[34] G. Ashton, N. Bernstein, J. Buchner, X. Chen, G. Csányi, A. Fowlie, F. Feroz, M. Griffiths, W. Handley, M. Habeck, E. Higson, M. Hobson, A. Lasenby, D. Parkinson, L. B. Pártay, M. Pitkin, D. Schneider, J. S. Speagle, L. South, J. Veitch, P. Wacker, D. J. Wales, and D. Yallup, Nature Reviews Methods Primers **2**, 39 (2022), arXiv:2205.15570 [stat.CO].

[35] I. Verdinelli and L. Wasserman, Journal of the American Statistical Association **90**, 614 (1995).

[36] S. R. Furlanetto, S. P. Oh, and F. H. Briggs, Phys. Rep. **433**, 181 (2006), arXiv:astro-ph/0608032 [astro-ph].

[37] J. D. Bowman, A. E. E. Rogers, R. A. Monsalve, T. J. Mozdzen, and N. Mahesh, Nature **555**, 67 (2018), arXiv:1810.05912 [astro-ph.CO].

[38] R. Hills, G. Kulkarni, P. D. Meerburg, and E. Puchwein, Nature **564**, E32 (2018), arXiv:1805.01421 [astro-ph.CO].

[39] R. Barkana, Nature **555**, 71 (2018), arXiv:1803.06698 [astro-ph.CO].

[40] P. H. Sims and J. C. Pober, MNRAS **492**, 22 (2020), arXiv:1910.03165 [astro-ph.CO].

[41] S. Singh, N. T. Jishnu, R. Subrahmanyan, N. Udaya Shankar, B. S. Girish, A. Raghunathan, R. Somashekar, K. S. Srivani, and M. Sathyanarayana Rao, Nature Astronomy **6**, 607 (2022), arXiv:2112.06778 [astro-ph.CO].

[42] L. Philip, Z. Abdurashidova, H. C. Chiang, N. Ghazi, A. Gumba, H. M. Heilgendorff, J. M. Jáuregui-García, K. Malepe, C. D. Nunhokee, J. Peterson, J. L. Sievers, V. Simes, and R. Spann, Journal of Astronomical Instrumentation **8**, 1950004 (2019).

[43] H. T. J. Bevins, W. J. Handley, A. Fialkov, E. de Lera Acedo, and K. Javid, MNRAS **508**, 2923 (2021), arXiv:2104.04336 [astro-ph.CO].

[44] E. Visbal, R. Barkana, A. Fialkov, D. Tseliakhovich, and C. M. Hirata, Nature **487**, 70 (2012), arXiv:1201.1005 [astro-ph.CO].

[45] A. Fialkov, R. Barkana, A. Pinhas, and E. Visbal, MNRAS **437**, L36 (2014), arXiv:1306.2354 [astro-ph.CO].

[46] I. Reis, A. Fialkov, and R. Barkana, MNRAS **499**, 5993 (2020), arXiv:2008.04315 [astro-ph.CO].

[47] Planck Collaboration, N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, R. Battye, K. Benabed, J. P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J. F. Cardoso, J. Carron, A. Challinor, H. C. Chiang, J. Chluba, L. P. L. Colombo, C. Combet, D. Contreras, B. P. Crill, F. Cuttaia, P. de Bernardis, G. de Zotti, J. Delabrouille, J. M. Delouis, E. Di Valentino, J. M. Diego, O. Doré, M. Douspis, A. Ducout, X. Dupac, S. Dusini, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, Y. Fantaye, M. Farhang, J. Fergusson, R. Fernandez-Cobos, F. Finelli, F. Forastieri, M. Frailis, A. A. Fraisse, E. Franceschi, A. Frolov, S. Galeotta, S. Galli, K. Ganga, R. T. Génova-Santos, M. Gerbino, T. Ghosh, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Handley, F. K. Hansen, D. Herranz, S. R. Hildebrandt, E. Hivon, Z. Huang, A. H. Jaffe, W. C. Jones, A. Karakci, E. Keihänen, R. Keskitalo, K. Kiiveri, J. Kim, T. S. Kisner, L. Knox, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, G. Lagache, J. M. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, P. Lemos, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Lilley, V. Lindholm, M. López-Caniego, P. M. Lubin, Y. Z. Ma, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marcos-Caballero, M. Maris, P. G. Martin, M. Martinelli, E. Martínez-González, S. Matarrese, N. Mauri, J. D. McEwen, P. R. Meinhold, A. Melchiorri, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M. A. Miville-Deschênes, D. Molinari, L. Montier, G. Morgante, A. Moss, P. Natoli, H. U. Nørgaard-Nielsen, L. Pagano, D. Paoletti, B. Partridge, G. Patanchon, H. V. Peiris, F. Perrotta, V. Pettorino, F. Piacentini, L. Polastri, G. Polenta, J. L. Puget, J. P. Rachen, M. Reinecke, M. Remazeilles, A. Renzi, G. Rocha, C. Rosset, G. Roudier, J. A. Rubiño-Martín, B. Ruiz-Granados, L. Salvati, M. Sandri, M. Savelainen, D. Scott, E. P. S. Shellard, C. Sirignano, G. Sirri, L. D.

Spencer, R. Sunyaev, A. S. Suur-Uski, J. A. Tauber, D. Tavagnacco, M. Tenti, L. Toffolatti, M. Tomasi, T. Trombetti, L. Valenziano, J. Valiviita, B. Van Tent, L. Vibert, P. Vielva, F. Villa, N. Vittorio, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Zacchei, and A. Zonca, A&A **641**, A6 (2020), arXiv:1807.06209 [astro-ph.CO].

[48] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems (2015), software available from tensorflow.org.

[49] A. Kessy, A. Lewin, and K. Strimmer, arXiv e-prints , arXiv:1512.00809 (2015), arXiv:1512.00809 [stat.ME].

[50] W. J. Handley, M. P. Hobson, and A. N. Lasenby, MNRAS **450**, L61 (2015), arXiv:1502.01856 [astro-ph.CO].

[51] W. J. Handley, M. P. Hobson, and A. N. Lasenby, MNRAS **453**, 4384 (2015), arXiv:1506.00171 [astro-ph.IM].

[52] Z. Abdurashidova, J. E. Aguirre, P. Alexander, Z. S. Ali, Y. Balfour, R. Barkana, A. P. Beardsley, G. Bernardi, T. S. Billings, J. D. Bowman, R. F. Bradley, P. Bull, J. Burba, S. Carey, C. L. Carilli, C. Cheng, D. R. DeBoer, M. Dexter, E. de Lera Acedo, J. S. Dillon, J. Ely, A. Ewall-Wice, N. Fagnoni, A. Fialkov, R. Fritz, S. R. Furlanetto, K. Gale-Sides, B. Glendenning, D. Gorthi, B. Greig, J. Grobbelaar, Z. Halday, B. J. Hazelton, S. Heimersheim, J. N. Hewitt, J. Hickish, D. C. Jacobs, A. Julius, N. S. Kern, J. Kerrigan, P. Kittiwisit, S. A. Kohn, M. Kolopanis, A. Lanman, P. La Plante, T. Lekalake, D. Lewis, A. Liu, Y.-Z. Ma, D. MacMahon, L. Malan, C. Malgas, M. Maree, Z. E. Martinot, E. Matsetela, A. Mesinger, J. Mirocha, M. Molewa, M. F. Morales, T. Mosiane, J. B. Muñoz, S. G. Murray, A. R. Neben, B. Nikolic, C. D. Nunhokee, A. R. Parsons, N. Patra, S. Pieterse, J. C. Pober, Y. Qin, N. Razavi-Ghods, I. Reis, J. Ringuette, J. Robnett, K. Rosie, M. G. Santos, S. Sikder, P. Sims, C. Smith, A. Syce, N. Thyagarajan, P. K. G. Williams, and H. Zheng, ApJ **924**, 51 (2022), arXiv:2108.07282 [astro-ph.CO].

[53] HERA Collaboration, Z. Abdurashidova, T. Adams, J. E. Aguirre, P. Alexander, Z. S. Ali, R. Baartman, Y. Balfour, R. Barkana, A. P. Beardsley, G. Bernardi, T. S. Billings, J. D. Bowman, R. F. Bradley, D. Breitman, P. Bull, J. Burba, S. Carey, C. L. Carilli, C. Cheng, S. Choudhuri, D. R. DeBoer, E. de Lera Acedo, M. Dexter, J. S. Dillon, J. Ely, A. Ewall-Wice, N. Fagnoni, A. Fialkov, R. Fritz, S. R. Furlanetto, K. Gale-Sides, H. Garsden, B. Glendenning, A. Gorce, D. Gorthi, B. Greig, J. Grobbelaar, Z. Halday, B. J. Hazelton, S. Heimersheim, J. N. Hewitt, J. Hickish, D. C. Jacobs, A. Julius, N. S. Kern, J. Kerrigan, P. Kittiwisit, S. A. Kohn, M. Kolopanis, A. Lanman, P. La Plante, D. Lewis, A. Liu, A. Loots, Y.-Z. Ma, D. H. E. MacMahon, L. Malan, K. Malgas, C. Malgas, M. Maree, B. Marero, Z. E. Martinot, L. McBride, A. Mesinger, J. Mirocha, M. Molewa, M. F. Morales, T. Mosiane, J. B. Muñoz, S. G. Murray, V. Nagpal, A. R. Neben, B. Nikolic, C. D. Nunhokee, H. Nuwegeld, A. R. Parsons, R. Pascua, N. Patra, S. Pieterse, Y. Qin, N. Razavi-Ghods, J. Robnett, K. Rosie, M. G. Santos, P. Sims, S. Singh, C. Smith, H. Swarts, J. Tan, N. Thyagarajan, M. J. Wilensky, P. K. G. Williams, P. van Wyngaarden, and H. Zheng, ApJ **945**, 124 (2023).

[54] H. T. J. Bevins, A. Fialkov, E. de Lera Acedo, W. J. Handley, S. Singh, R. Subrahmanyan, and R. Barkana, Nature Astronomy **6**, 1473 (2022), arXiv:2212.00464 [astro-ph.CO].

[55] M. Evans, A. Corsi, C. Afle, A. Ananyeva, K. G. Arun, S. Ballmer, A. Bandopadhyay, L. Barsotti, M. Baryakhtar, E. Berger, E. Berti, S. Biscoveanu, S. Borhanian, F. Broekgaarden, D. A. Brown, C. Cahillane, L. Campbell, H.-Y. Chen, K. J. Daniel, A. Dhani, J. C. Driggers, A. Effler, R. Eisenstein, S. Fairhurst, J. Feicht, P. Fritschel, P. Fulda, I. Gupta, E. D. Hall, G. Hammond, O. A. Hannuksela, H. Hansen, C.-J. Haster, K. Kacanja, B. Kamai, R. Kashyap, J. Shapiro Key, S. Khadkikar, A. Kontos, K. Kuns, M. Landry, P. Landry, B. Lantz, T. G. F. Li, G. Lovelace, V. Mandic, G. L. Mansell, D. Martynov, L. McCuller, A. L. Miller, A. H. Nitz, B. J. Owen, C. Palomba, J. Read, H. Phurailatpam, S. Reddy, J. Richardson, J. Rollins, J. D. Romano, B. S. Sathyaprakash, R. Schofield, D. H. Shoemaker, D. Sigg, D. Singh, B. Slagmolen, P. Sledge, J. Smith, M. Soares-Santos, A. Strunk, L. Sun, D. Tanner, L. A. C. van Son, S. Vitale, B. Willke, H. Yamamoto, and M. Zucker, arXiv e-prints , arXiv:2306.13745 (2023), arXiv:2306.13745 [astro-ph.IM].

[56] M. Milgrom, ApJ **270**, 365 (1983).

[57] R. M. Wald, *General Relativity* (University of Chicago Press, 1984).