

Calibrating Bayesian tension statistics using neural ratio estimators

Harry T. J. Bevins^{1,2,*} William J. Handley^{2,3} and Thomas Gessey-Jones^{1,2}

¹*Astrophysics Group, Cavendish Laboratory, Cambridge, CB3 0HE, United Kingdom*

²*Kavli Institute for Cosmology, Cambridge, CB3 0HA, United Kingdom*

³*Institute of Astronomy, Cambridge, CB3 0HA, United Kingdom*



(Received 22 July 2024; accepted 21 July 2025; published 15 August 2025)

When fits of the same physical model to two different datasets disagree, we call this tension. Several apparent tensions in cosmology have occupied researchers in recent years, and a number of different metrics have been proposed to quantify tension. Many of these metrics suffer from limiting assumptions, and correctly calibrating these is essential if we want to successfully determine whether discrepancies are significant. A commonly used metric of tension is the evidence ratio R . The statistic has been widely adopted by the community as a Bayesian way of quantifying tensions, however, it has a nontrivial dependence on the prior that is not always accounted for properly. We show that this can be calibrated out effectively with neural ratio estimation. We demonstrate our proposed calibration technique with an analytic example, a toy example inspired by 21-cm cosmology, and with observations of the baryon acoustic oscillations from the Dark Energy Spectroscopic Instrument (DESI) and the Sloan Digital Sky Survey (SDSS). We find no significant tension between DESI and SDSS.

DOI: [10.1103/47hn-h8kc](https://doi.org/10.1103/47hn-h8kc)

I. INTRODUCTION

Independently confirming conclusions about the nature of our Universe from one experiment with another is crucial to the advance of knowledge. When the inference from two different experiments disagree with each other, we call this tension. Tension between different datasets raises questions about the need for new physics and better descriptions of our instruments and systematics. The H_0 and σ_8 tensions e.g. [1–5] are the most commonly encountered examples in cosmology, but other examples include observations of the 21-cm signal from cosmic dawn [6], tensions in the amplitude of the matter power spectrum [7] and tension in estimates of the curvature of the Universe [8]. A historical example of tension is in the measurement of the matter density Ω_m [9–12] which was resolved by the discovery of the accelerating universe [13]. A detailed review of cosmological tensions can be found in [14].

Many different measures of tension have been proposed and are widely used in cosmological studies. Examples include Bayesian suspiciousness [15], estimators of the probability of observed parameter differences [16,17], goodness of fit degradation [18] and Eigentension [19].

These tension statistics are summarized and reviewed in [20–22]. It is common practice to rephrase these tension metrics in σ units of tension, corresponding to probabilities on a one dimensional normal distribution. When expressed in this way, one would expect that the various tension metrics predict the same level of tension or concordance between different datasets. However, this is often not the case due to the various assumptions that are made when defining the statistics. To tackle this issue, we can try to define tension metrics that do not make these assumptions, however, the tension statistics often lose some of their interpretability when we do this. Instead, we try to calibrate out these assumptions in sensible ways. Calibration of tension statistics is an important and often overlooked step that needs to be taken to correctly interrogate the tension between different datasets.

A commonly used metric of tension is R corresponding to the ratio of a joint evidence and the product of individual evidences for two different datasets under a common model. The R statistic was first proposed in [23] and has been used to quantify tension in a number of cosmological studies e.g. [24,25]. R has also been used to perform model comparison in some works, although the authors of [26] showed that this approach to model comparison is incomplete.

The ratio R suffers from a nontrivial dependence on the prior that is not always accounted for properly [15]. In [15] the authors showed that as one decreases the prior width on the common parameters in the model of the two datasets, then the tension between the datasets should increase and R should decrease. Intuitively, one can see that it is more

*Contact author: htjb2@cam.ac.uk

satisfying if the two experiments favor parameters that are close together given a wide prior in comparison to a narrow prior. However, what constitutes “narrow” and “wide” is problem specific and subjective, making the interpretation of R difficult. We would like to calibrate out the prior dependence.

The authors of [15] propose an alternative statistic that is closely related to R called the suspiciousness S . Suspiciousness is calculated by subtracting a correction from R that quantifies the probability that the datasets would match given the prior ranges on the model parameters. It can be thought of as the value of R in the case where the priors are as narrow as possible without impacting the shape of the posterior. In [27] the authors convert S into σ s of tension, however, this requires an estimate of the number of constrained dimensions d in the joint analysis. To estimate d they use the Bayesian (sometimes referred to as Gaussian) model dimensionality, however, this is a poor estimator of d if the posterior is significantly non-Gaussian, as is often the case in cosmology.

In [21] the authors demonstrate that simulations can be used to calibrate tension metrics with the Planck data and Dark Energy Survey (DES) data. They proposed taking a fiducial set of parameters, such as the maximum posterior point for Planck, shifting these parameter values by some posterior-informed step sizes to induce a known degree of tension, simulate the now in tension DES observations and calculate the value of one’s chosen tension metric between the real Planck data and the simulation. To do this, however, one often has to run expensive sampling algorithms on the simulated data to calculate tension statistics such as R and S .

We propose calibrating the prior dependence of R using neural ratio estimation (NRE) e.g. [28–30]. NREs are classifiers, with interpretable outputs, that determine whether two quantities are drawn from independent distributions or a joint distribution. We show that the output of an NRE trained on simulations of two experiments’ observables can be interpreted as R and that an appropriately trained NRE can be used to calibrate for the prior dependence of the dimensionless R statistic. Since R requires the calculation of three Bayesian evidences, it is an expensive statistic to evaluate using traditional methods like nested sampling. We show that R can be accessed with a significantly smaller computational overhead using cutting edge machine learning tools. We call our NRE setup the TENSIONNET.

In Sec. II we summarize Bayesian inference and give more details about R . In Sec. III we discuss the interpretation of R and follow this with a discussion on NREs in Sec. IV. We discuss calibrating R with NREs in Sec. V. We then test our method on toy examples with known concordance R distributions in Sec. VI. We then apply the TENSIONNET to a toy example inspired by 21-cm cosmology and to assess the tension between baryon acoustic oscillations (BAO) observations from the Dark Energy

Spectroscopic Instrument (DESI) and the Sloan Digital Sky Survey (SDSS) in Sec. VII. We consider some limitations of our method in Sec. VIII and conclude in Sec. IX.

The code used in this paper is publicly available at¹.

II. BAYESIAN INFERENCE AND TENSION STATISTICS

In Bayesian inference we are interested in modeling data D with a model M containing parameters θ to recover both the probability of the data given the model $\mathcal{Z} = P(D|M)$ or evidence and the probability of a given θ given the data and model $P(\theta) = P(\theta|D, M)$ or posterior. To do this we draw samples from a prior $\pi(\theta) = P(\theta|M)$ which encodes our prior knowledge of the parameters and evaluate a likelihood which is our postulated probability of the data given a set of parameters and model $L(\theta) = P(D|\theta, M)$. We relate these quantities using Bayes theorem

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

where $\mathcal{Z}$ is given by

$$\mathcal{Z} = \int \mathcal{L}(\theta)\pi(\theta)d\theta. \quad (2)$$

An efficient and accurate way to recover both the evidence and the posterior is with nested sampling [31,32] although other methods exist e.g. [33–37].

The tension R between two datasets, indicated by the subscripts A and B , is

$$R = \frac{\mathcal{Z}_{A,B}}{\mathcal{Z}_A \mathcal{Z}_B} = \frac{P(D_A, D_B|M)}{P(D_A|M)P(D_B|M)} = \frac{P(D_A, D_B)}{P(D_A)P(D_B)}, \quad (3)$$

where we have dropped the dependence on M in the last expression for conciseness. R is prior dependent and this can be seen by noting that

$$\begin{aligned} R &= \frac{\mathcal{Z}_{A,B}}{\mathcal{Z}_A \mathcal{Z}_B} = \frac{1}{\mathcal{Z}_A \mathcal{Z}_B} \int \mathcal{L}_A \mathcal{L}_B \pi d\theta \\ &= \int \frac{\mathcal{L}_A \pi}{\mathcal{Z}_A} \frac{\mathcal{L}_B \pi}{\mathcal{Z}_B} \frac{1}{\pi} d\theta = \int \frac{P_A P_B}{\pi} d\theta \\ &= \left\langle \frac{P_B}{\pi} \right\rangle_{P_A} = \left\langle \frac{P_A}{\pi} \right\rangle_{P_B} \end{aligned} \quad (4)$$

where we have assumed the datasets are independent, and the angled brackets represent averages over the distributions P_A and P_B [15]. For a uniform prior, $\pi = 1/V$ where V is the volume, one can see from Eq. (4) that if the prior is made smaller than R being proportional to V also

¹<https://github.com/htjb/tension-networks>

decreases. This logic generalizes to more complicated priors.

III. INTERPRETING R

R has several attractive properties [15];

- (i) Dimensionally consistent: Whilst the Bayesian evidence typically carries units that are the inverse of the measure on the data space, by taking the ratio of evidences the units cancel. R is therefore not dependent on the units of the model space.
- (ii) Parametrization invariant: The value of R does not change if you reparametrize the model (e.g. switch from mass to log mass). While priors and likelihoods can change under reparametrizations the Bayesian evidence should properly account for this and ratios of evidences will remain invariant to these changes. R therefore reflects the true model-data compatibility and is not sensitive to arbitrary coordinate choices.
- (iii) Symmetric: Swapping the two datasets does not affect the value of R . So there is no preferred ordering and both datasets are treated equally.

It is typically interpreted with respect to a value of 1 with $R \ll 1$ corresponding to inconsistent datasets and $R \gg 1$ to consistent data. However, this interpretation does not tell us the degree to which our datasets are in tension given the prior and model choice. To try to quantify the tension between observations from the Dark Energy Survey and Planck, the authors of [38] interpreted R on a Jefferys' scale [39]. The Jefferys' scale is, however, somewhat arbitrary.

In [15] the authors showed that

$$R = \frac{\mathcal{Z}_{A,B}}{\mathcal{Z}_A \mathcal{Z}_B} = \frac{P(D_A, D_B)}{P(D_A)P(D_B)} = \frac{P(D_A|D_B)}{P(D_A)} = \frac{P(D_B|D_A)}{P(D_B)}, \quad (5)$$

implying then one can interpret R , if it is greater than 1, as a fractional increase in confidence in dataset A given knowledge of dataset B over A alone (or vice versa). If $R \ll 1$ then the authors suggest we should be concerned about our model or the datasets.

When interpreting R one has to keep in mind the impact which the prior has on its value. Reducing the width of the prior will increase the apparent tension between the datasets by reducing the value of R . The authors of [15] suggest repeating our analysis with sensible modifications to the prior distribution to determine how stable the value of R is and consequently our conclusions regarding the tension between different datasets.

Given a choice of prior and model, there is a distribution of possible in concordance R values that could be observed between two different experiments. Low signal-to-noise observations of the same signal by the two experiments will have lower typical values of R in contrast to high signal-to-noise observations. This distribution can be used to

translate between R and $N\sigma$ estimates of tension, removing the prior dependence from the statistic and allowing for comparison with other tension metrics. The difficulty, however, is in accessing this distribution, which requires the evaluation of individual and joint evidences for a large sample of simulations, making it computationally expensive and often intractable. In this paper, we propose calibrating the prior dependence of R using simulations and NREs to quickly evaluate the in concordance R distribution.

IV. NEURAL RATIO ESTIMATION

NREs are neural network classifiers that are trained to return the probability that two inputs have been drawn from a joint distribution relative to the probability that they have been drawn from independent distributions. For training data that includes an equal number of examples of two inputs A and B drawn from their independent distributions and their joint distribution, the output of a neural ratio estimator tends toward

$$\log r = \log \frac{P(A, B)}{P(A)P(B)}. \quad (6)$$

To prove this, we begin by defining the network output as $f(A, B)$. During training, we give it examples drawn from the joint distribution $P(A, B)$ with probability P_J and drawn from $P(A)P(B)$ with probability $(1 - P_J)$. NREs are trained with a binary cross entropy loss function that is defined as

$$l = \frac{1}{N} \left[\sum_i y_i \log(\tilde{f}(A, B)) + (1 - y_i) \log(1 - \tilde{f}(A, B)) \right], \quad (7)$$

where

$$\tilde{f}(A, B) \equiv S_\sigma(f(A, B)) = \frac{e^{f(A, B)}}{1 + e^{f(A, B)}}, \quad (8)$$

and where y_i is 1 for samples drawn from the joint and 0 for independent samples. S_σ is the sigmoid activation function and scales the output of the network between 0 and 1.

In the limit of a large number of training samples, we can take the continuous limit of the sum

$$l \approx - \int P(A, B) P_J \log(\tilde{f}(A, B)) + P(A)P(B)(1 - P_J) \log(1 - \tilde{f}(A, B)) dA dB, \quad (9)$$

where the approximation approaches equality as the size of the training dataset approaches infinity. During training the loss function is minimized and so we can find the function

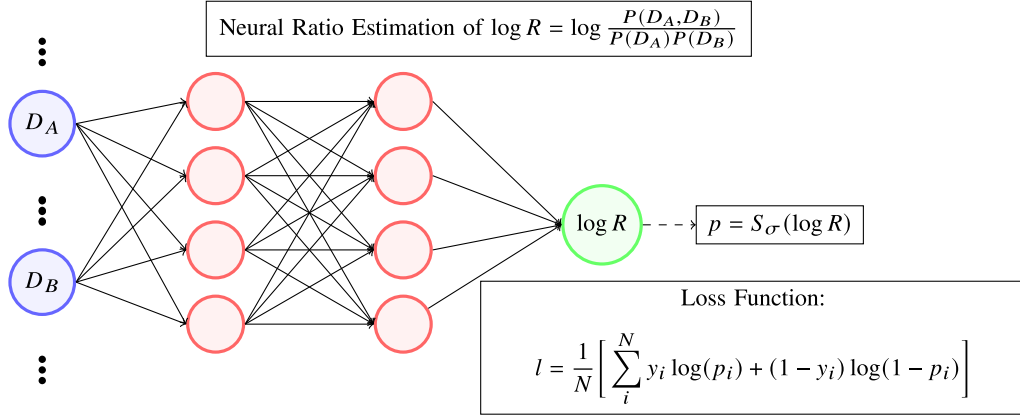


FIG. 1. Schematic of the NRE used in this work, which we refer to as a TENSIONNET. The NRE is trained on matched and mismatched pairs of simulated observations from two different experiments A and B and outputs an estimate of the tension statistic R . The network is trained using the binary cross entropy loss function.

the network should converge to via the calculus of variations

$$0 = \frac{\delta l}{\delta \tilde{f}} = \frac{P(A, B)P_J}{\tilde{f}(A, B)} - \frac{P(A)P(B)(1 - P_J)}{1 - \tilde{f}(A, B)}, \quad (10)$$

which can be rewritten as

$$\tilde{f}(A, B) = \frac{\frac{P(A, B)P_J}{P(A)P(B)(1 - P_J)}}{1 + \frac{P(A, B)P_J}{P(A)P(B)(1 - P_J)}}. \quad (11)$$

Recalling that the output of our network is defined such that $\tilde{f}(A, B) = S_\sigma(f(A, B))$ we see that

$$f(A, B) \rightarrow \log \left(\frac{P(A, B)P_J}{P(A)P(B)(1 - P_J)} \right), \quad (12)$$

which when $P_J = 0.5$ gives

$$f(A, B) \rightarrow \log r, \quad (13)$$

where in the limit of perfect training $f(A, B) = \log r$.

V. CALIBRATING R WITH NRES

As discussed above, a trained NRE outputs the log of the ratio

$$r = \frac{P(A, B)}{P(A)P(B)}. \quad (14)$$

It can be seen, trivially,

$$r = R = \frac{P(D_A, D_B)}{P(D_A)P(D_B)} = \frac{\mathcal{Z}_{A,B}}{\mathcal{Z}_A \mathcal{Z}_B}, \quad (15)$$

if the inputs to the NRE A and B correspond to the datasets D_A and D_B .

We propose that the true observed tension R_{obs} is calculated using nested sampling e.g. [15] or an alternative independent evidence estimation tool. Then we propose using the NRE to predict the in concordance R distribution, against which one can calibrate R_{obs} . A schematic of the NRE or TENSIONNET is shown in Fig. 1.

In practice, our proposed calibration method is as follows;

- (1) Generate a set of matched simulations, using the same models and prior used to evaluate R_{obs} , of $D_A(\theta)$ and $D_B(\theta)$ where they share the same parameters. This gives us the set $s = \{D_A(\theta_i), D_B(\theta_i)\}_{i=0}^N$.
- (2) We then shuffle one set of the simulations to give us $s' = \{D_A(\theta_i), D_B(\theta_j)\}_{i \neq j=0}^N$.
- (3) We label the matched sets of data with a value of 1 and the mismatched data with a value of 0.
- (4) We then shuffle our labeled matched and mismatched data and split this into training and validation data.
- (5) We then train our neural ratio estimator and perform early stopping using the validation data.
- (6) Once trained we then generate a new set of matched datasets, $z = \{D_A(\theta_i), D_B(\theta_i)\}_{i=0}^N$, from the models covering the entire prior range and calculate their corresponding $\log R$ values with the NRE to recover the in concordance distribution.
- (7) Given samples on this distribution $P(\log R)$ we then calculate an empirical cumulative distribution function (CDF), $P(\log R < \log R')$ which along with the inverse survival function of the standard normal distribution can be used to translate R into the desired prior calibrated $N\sigma$ measure of tension.

The inverse survival function $z(\alpha)$ is defined as the probability that a random variable X takes a value less than x . Specifically, we are interested in the one-sided

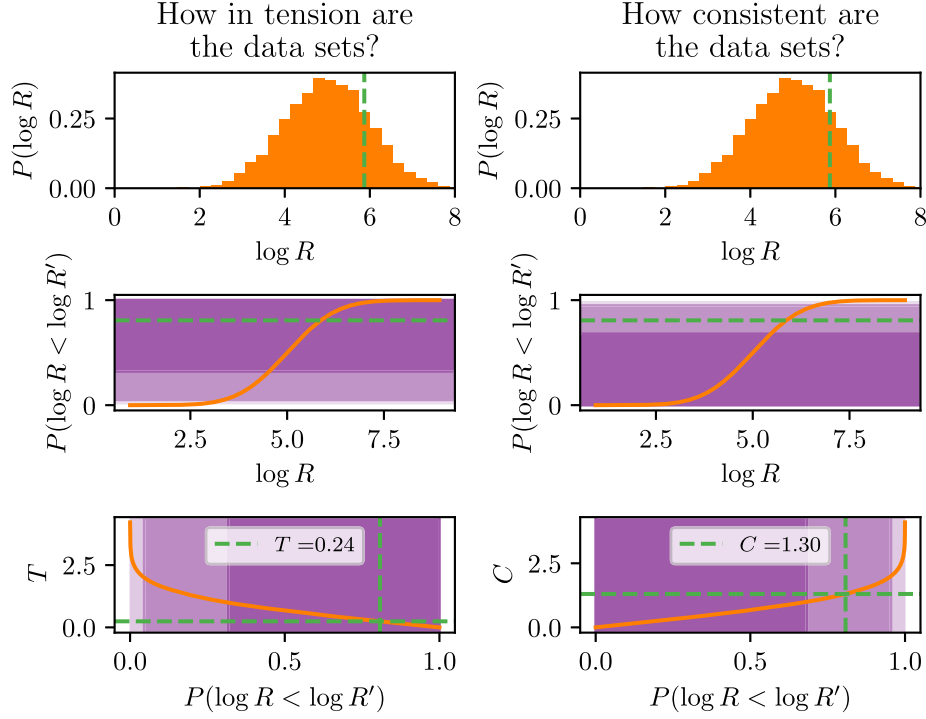


FIG. 2. Interpreting R_{obs} with NREs. The top row of the figure shows an example distribution of possible in concordance R values. As we move to the right of the median of the distribution we move toward concordance and to the left, lower values of $\log R$, toward tension. The middle row of the figure shows the corresponding cumulative distribution function, and the bottom row shows how the tension statistic T and concordance statistic C vary with $\log R$ for this example. The observed $\log R_{\text{obs}}$, its corresponding value on the CDF and its value on T and C are shown as green dashed lines. The shaded regions show the 1, 2 and 3σ contours for both statistics with the darker region representing 1σ and the lighter region 3σ .

inverse survival function which for a standard normal distribution is

$$z\left(\frac{\alpha}{2}\right) = \sqrt{2}\text{erf}^{-1}\left(2\left(1 - \frac{\alpha}{2}\right) - 1\right). \quad (16)$$

We can define a prior calibrated tension statistic from the CDF of the $\log R$ distribution

$$\begin{aligned} T &= z\left(\frac{P(\log R < \log R')}{2}\right) \\ &= \sqrt{2}\text{erf}^{-1}(1 - P(\log R < \log R')). \end{aligned} \quad (17)$$

If $P(\log R < \log R') = 1$ then $T = 0$ and we should be concerned that are datasets are in perfect agreement. If $T = 3$ for example, then we can say that the experiments are in 3σ tension. Conversely, we can define a concordance statistic

$$\begin{aligned} C &= z\left(\frac{(1 - P(\log R < \log R'))}{2}\right) \\ &= \sqrt{2}\text{erf}^{-1}(P(\log R < \log R')), \end{aligned} \quad (18)$$

where a value of $C = 3$ indicates a 3σ agreement between the datasets. If C becomes very large, then we would

conclude that the datasets are in a suspiciously high degree of agreement (see Fig. 2).

VI. VALIDATING THE NRE

To demonstrate the robustness of our method, and some of its limitations, we first look at an example with an analytically tractable distribution of in concordance $\log R$ and compare this with the prediction from the NRE.

We begin by defining our prior and likelihood function in our example to be Gaussian and use a linear model for each of our observed datasets,

$$\begin{aligned} D &= M\theta + m \pm \sqrt{C}, \\ \mathcal{L}(D|\theta) &= \mathcal{N}(M\theta + m, C), \\ \pi(\theta) &= \mathcal{N}(\mu, \Sigma) \end{aligned} \quad (19)$$

where θ are the model parameters, M and m define the data model and data samples can be drawn from the likelihood with covariance C . In the example that follows M , m and C are different for each experiment. μ and Σ are the mean and covariance of our prior. In such a setup the Bayesian evidence for each experiment and the joint observation are analytically tractable. For each experiment, the evidence is given by

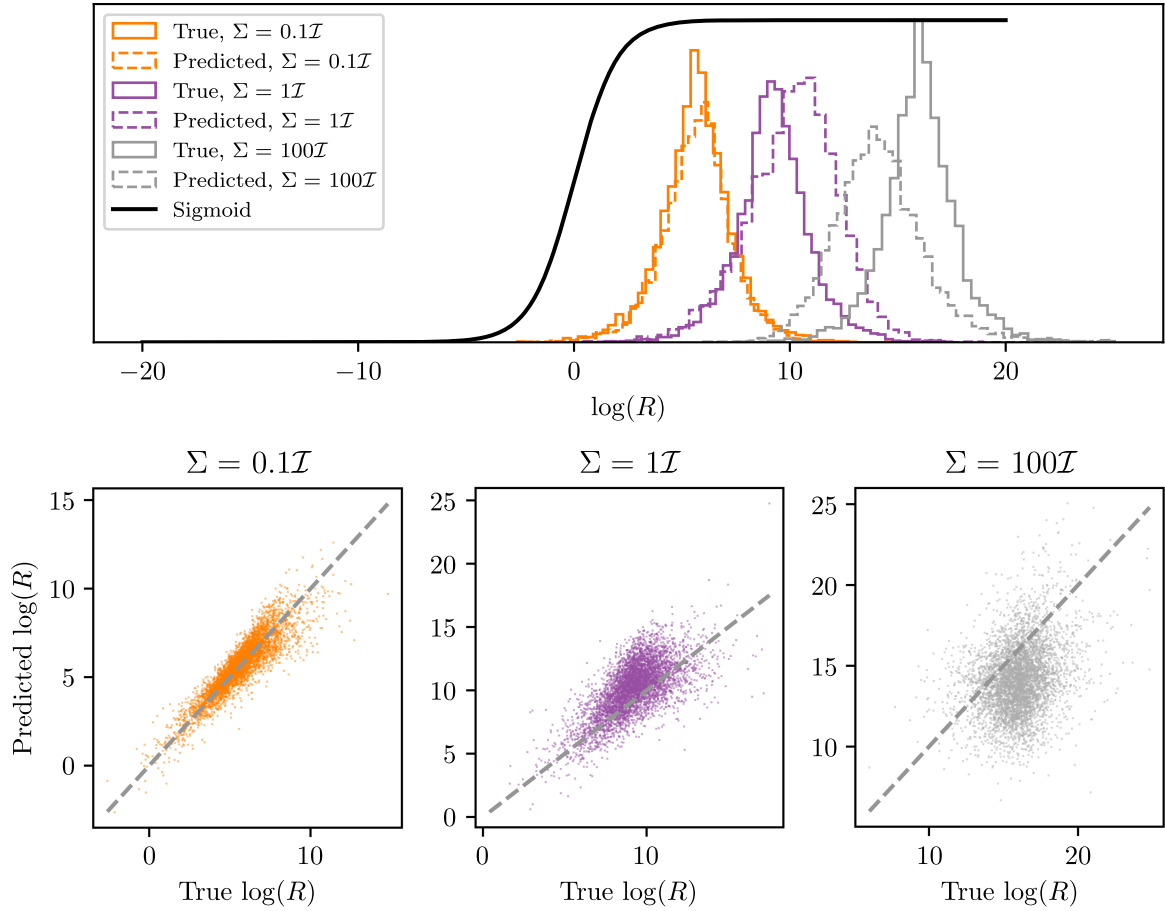


FIG. 3. We hypothesize two experiments observing data that can be described with a linear model and a Gaussian likelihood function. By then defining our prior to also be Gaussian with a diagonal covariance Σ we can analytically calculate the joint and individual evidences and the tension statistic R . We draw a test set from the joint distribution $\mathcal{Z}(D_A, D_B)$ which we use to analytically derive the in concordance $\log R$ distribution (solid lines, top panel) and predict the distribution from the NRE (dashed lines, top panel) for different prior widths. We also show the sigmoid activation function for reference. The bottom row shows the predicted versus true $\log R$ values for the test set for different prior widths. Performance begins to break down for $\log R > 10$.

$$\mathcal{Z} = \mathcal{N}(m + M\mu, C + M\Sigma M^T), \quad (20)$$

where M^T is the transpose of M . We use the LSBI package to evaluate these expressions [40].

We draw training data from $\mathcal{Z}_{AB} = P(D_A, D_B)$ for our NRE and for each pair of D_A and D_B in the test data we analytically calculate R to build the “true” distribution that we are trying to predict with the trained NRE.

A. Assessing the performance of the NRE

The performance of NREs is known to degrade as the absolute value of the log ratio they are predicting increases. Therefore, we might expect the performance of the TENSIONNET to degrade with increasing prior width, and we test this by comparing the prediction from the NRE with the true distribution for a range of prior widths Σ .

We define M to be a matrix of uniform random numbers between 0 and 1 of dimensions $d \times n$ where $d = 50$ is the number of data points and $n = 3$ is the number of

dimensions. m and μ are defined to be a vector of uniform random numbers between 0 and 1 of length d and C is a diagonal matrix of 0.01. Where M and m vary, the prior defined by Σ and μ is the same for both experiments. Σ is a diagonal matrix, and we consider three different scenarios where $\Sigma = 0.1I, 1I$ and $100I$ where I is the identity matrix.

For each Σ we generate 500,000 matched observations from experiment A and experiment B for training the NRE. We use an exponentially decaying learning rate with an initial value of 10^{-3} , a step size of 1000 epochs and a decay rate of 0.9. We use a ReLU activation function in the hidden layers, five hidden layers of 25 nodes each, a maximum number of epochs of 1000 with early stopping and a batch size of 1000. We use the ADAM optimizer for training. Once trained, we generate a new set of 5000 previously unseen in concordance observations from the models to put through the NRE and generate a predicted distribution of $\log R$.

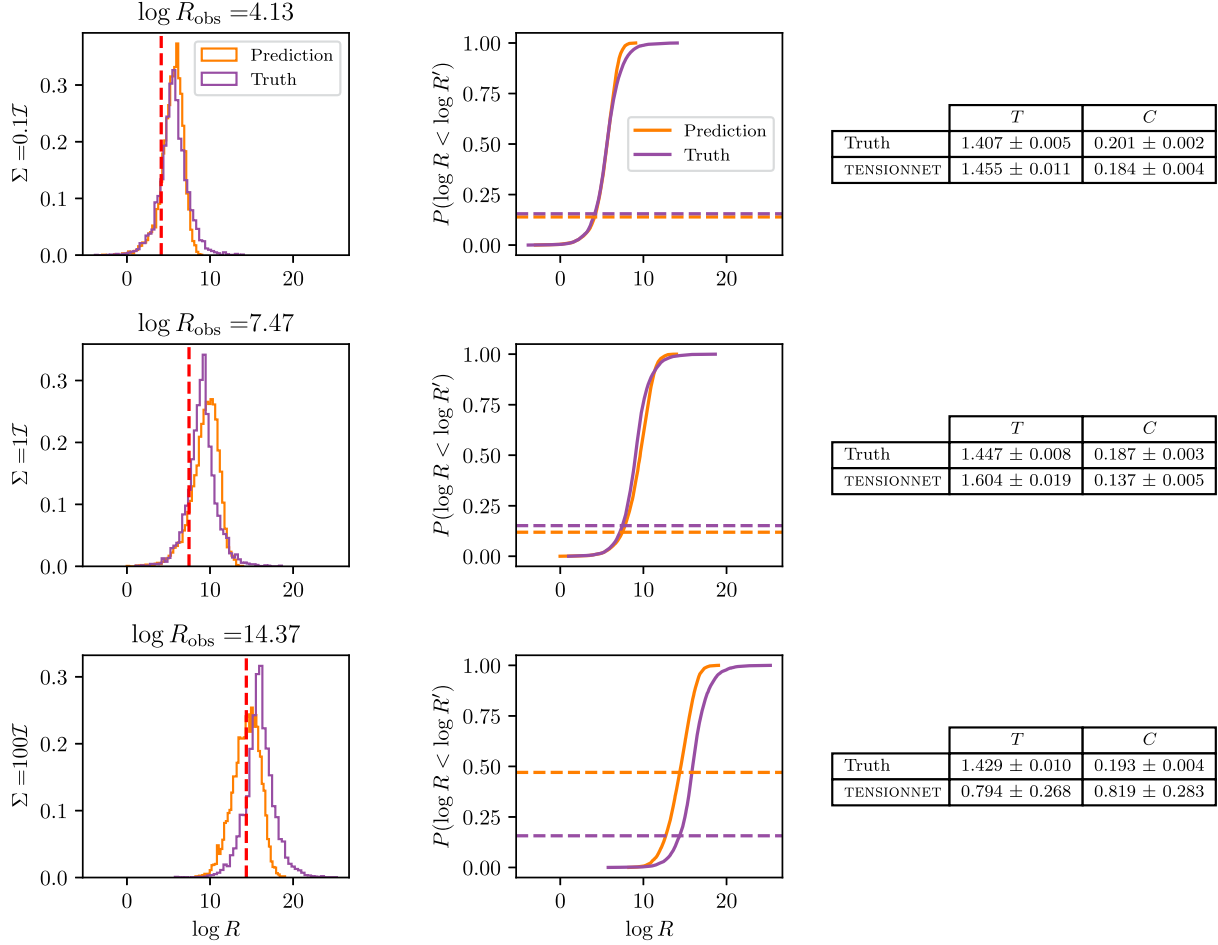


FIG. 4. Using the linear model described in Sec. VI we show how the in concordance $\log R$ distribution can be used to calibrate the prior dependence of the R statistic. We also show how the predicted in concordance R distribution from the TENSIONNET is largely consistent with the analytic distribution. The narrowest prior is on the top row and the widest on the bottom row. The first column shows the distribution of in concordance $\log R$ values calculated analytically in purple and as predicted by the NRE in orange. We also show the analytically calculated value of $\log R$ for a simulation drawn from the narrow prior as a red dashed line. The middle column shows the CDFs derived from the two in concordance distributions and as horizontal dashed lines the value of the CDF at R_{obs} according to the analytic (purple) and NRE (orange) distributions. The final column shows the average values over five runs of T and C derived using the true analytic distributions and the NRE for each prior with an associated error.

The top panel of Fig. 3 shows the predicted distributions (dashed lines) from the NRE versus the analytic distributions (solid lines) for different Σ . The solid black line shows the sigmoid activation function. The bottom three panels show the predicted versus true $\log R$ for each pair of data samples in the distribution. As the prior widens and $\log R$ becomes larger, the accuracy with which the distribution is recovered degrades as expected. Performance drops off, particularly for large prior widths, when $\log R > 10$. Caution needs to therefore be taken when using the NRE to calibrate values of $\log R \gg 10$. In such circumstances, one could consider reducing the width of the prior or running nested sampling on a handful of simulations to gauge how well the NRE is performing. If all one is interested in is the tension between different datasets, one could also choose one's prior so that $\log R$ is closer to 1 since the proposed tension metrics T and C are prior

independent. For the orange distribution with $\Sigma = 0.1\mathcal{L}$ and to some extent the purple distribution with $\Sigma = 1\mathcal{L}$, the NRE accurately recovers the $\log R$ distribution.

B. Calibrating out the prior

Using the above example, we can also illustrate how the TENSIONNET can be used to calibrate out the dependence of R on the prior. In Fig. 4, we keep our data model the same but change the prior width on our three parameters. Our observed dataset is drawn from the narrowest prior and kept the same throughout. We can clearly see that as the prior width increases, so does R_{obs} as expected. However, we can also see that the true distribution (purple) of in concordance $\log R$ values also shifts to higher values. When we use this distribution to calibrate $\log R_{\text{obs}}$ into T and C the values are approximately constant regardless of the prior width.

Calibrating against the predicted distribution from the NRE (orange) gives largely consistent results with some degradation in performance for the largest prior width as expected from the last section. We repeat the analysis five times and report the average values of T and C with an associated error for both the true and predicted distributions in Fig. 4.

VII. COSMOLOGICAL EXAMPLES

A. Toy 21-cm cosmology

Observers in the field of 21-cm cosmology are aiming to detect an information rich redshifted signal from neutral hydrogen from the Cosmic Dawn and Epoch of Reionization, see [41–44] for reviews of the field. The signal is observed in the radio band, and can in theory be detected as a sky-averaged 21-cm signal e.g. [45–47]. It has a complex dependence on the astrophysics of the early Universe e.g. [48–57] and can be thought of as a spectral distortion in the cosmic microwave background (CMB). The key challenge in 21-cm cosmology is the separation of this signal from the dominant Galactic and extragalactic foregrounds, that the instruments also observe, whilst accounting for the nonuniform response of the instruments to the sky [58].

Since we are focused on illustrating the performance of the TENSIONNET, we ignore the effects of foregrounds and the instrument in our example and use a simplistic Gaussian absorption feature to model the 21-cm signal. We include Gaussian distributed noise in our simulated data (inspired by current observations [45,46]) and a Gaussian absorption feature

$$\delta T_b = -A \exp\left(-\frac{(\nu - \nu_0)^2}{w^2}\right), \quad (21)$$

where A corresponds to the amplitude of the signal, ν_0 to the central frequency and w to the width. A number of works in the literature have adopted this model for illustrative purposes e.g. [58,59] but we stress that it is simplistic and does not capture the complex astrophysics encoded by the 21-cm signal.

Current observations of the sky-averaged 21-cm signal include a tentative detection by the EDGES Collaboration [45] and an upper limit on the magnitude of the signal from SARAS3 [46]. Analysis by the SARAS3 team suggested that these measurements are in tension with each other, and a number of works have discussed the possible presence of systematics in the EDGES data [59–62]. As more experiments come online in the coming years e.g. [47] the assessment of tension and concordance between different observations is going to become crucial for the field.

We simulate observations of the sky-averaged 21-cm signal from two different experiments in different frequency ranges. We hypothesize that the 21-cm signal has a depth of 0.2 K, a central frequency of 78 MHz

and a width of 10 MHz. In our example, the first experiment (Exp. A) has made a detection of the signal with Gaussian distributed noise with a standard deviation of 25 mK over the frequency range 60–90 MHz with a channel width of ≈ 0.3 MHz (see top left panel of Fig. 5). We then hypothesize a series of scenarios where a second experiment (Exp. B) has observed the 21-cm signal in the frequency range 80–120 MHz with a channel width of ≈ 0.4 MHz with the same central frequency and width but a different magnitude $A = [0.15, 0.2, 0.25]$ K such that the observations are in tension, concordance and tension, respectively. We add 25 mK Gaussian random noise to the data from experiment B.

We fit each pair of observations using the nested sampling implementation POLYCHORD [63,64] to assess R_{obs} . We use Eq. (21) as our model, M for the data, D and use a Gaussian likelihood function

$$\log \mathcal{L} = \sum_i -\frac{1}{2} \log 2\pi\sigma^2 - \frac{1}{2} \frac{(D_i - M_i)^2}{\sigma^2}, \quad (22)$$

where the sum is over observation frequency and σ is the standard deviation of the noise, which we fit as a free parameter. The prior is uniform on A between 0.0–4.0 K, ν_0 between 60–80 MHz, w between 5–40 MHz and σ between 0.001–0.1 K.

The combination of our likelihood and prior and the fact that our model is nonlinear makes the in concordance $\log R$ distribution analytically intractable. It can only be accessed in a reasonable amount of time through the TENSIONNET. One could of course evaluate the distribution with 1000s of Nested Sampling runs, but this would be computationally expensive. To calculate R using Nested Sampling for this example costs around 12 seconds (all the analysis performed in the paper was done on an M2 MacBook with 8 GB ram). This includes the run time of the inference on each dataset individually and on the joint dataset. Using Nested Sampling to evaluate R for 5000 pairs of simulated in concordance observations to assess the distribution of possible values would therefore take around 16.7 hrs. To generate simulations for the NRE, train the NRE and evaluate it for 5000 pairs of datasets to get the in concordance R distribution takes around 6.6 minutes. For this problem, the NRE is therefore $(16.7 \times 60)/6.6 \approx 152$ times faster. Note that this is a trivial problem with a fast likelihood function.

We generate 200,000 mock observations of the 21-cm signal for both experiments with the same sets of parameters from the prior. We then shuffle these datasets to create a corresponding set of in tension “observations” giving us a set of 400,000 simulations. We use 80% of this to train the NRE and the rest to perform early stopping.

Once trained, we generate 5000 pairs of observations of the same signal by both experiments with parameters drawn randomly from the prior range to evaluate the in

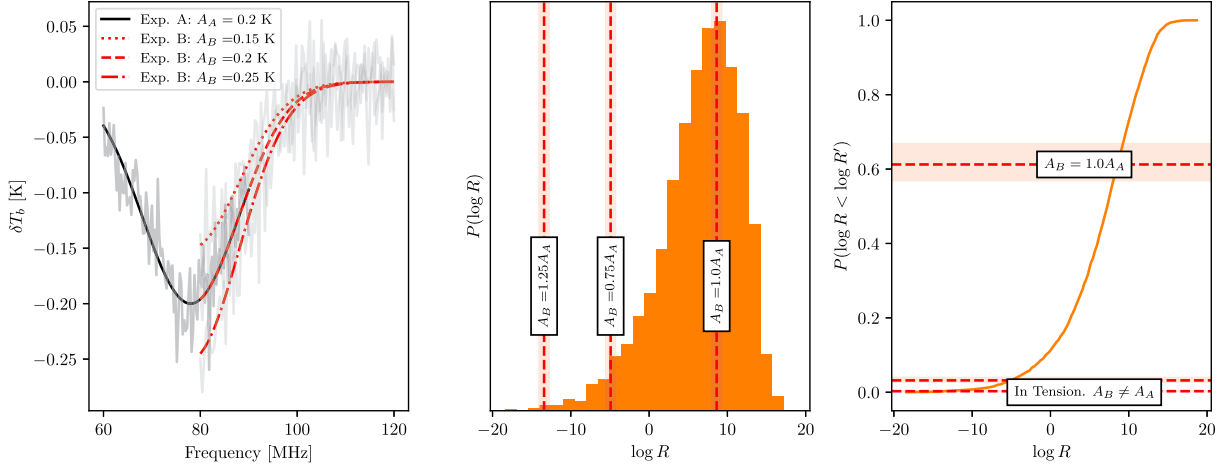


FIG. 5. To further illustrate the application of NREs to the calibration of R we use a toy example inspired by 21-cm cosmology. Left panel: We simulate an experiment observing a Gaussian absorption trough as a function of frequency (black line) and three different scenarios in which another experiment measures a 21-cm signal with either the same or different amplitudes in a different band (red lines). To each observation, we add Gaussian random noise with a standard deviation of 25 mK (shown in grey and motivated by current observations e.g. [45]). Middle panel: We train the NRE on simulated observations of the signal by both experiments, covering a wide prior range of signal parameters. We use the NRE to evaluate the possible distribution of in concordance $\log R$ values. We plot the observed $\log R$ for each pair of observations from experiments A and B. Right panel: Finally, we show the CDF of the in concordance $\log R$ distribution in the right panel of the figure and the corresponding CDF values for each pair of observations. We find that for the two in tension observations the $T = 2.989^{+0.167}_{-0.060}$ and $T = 2.147^{+0.056}_{-0.089}$ and for the in concordance observations $C = 0.864^{+0.107}_{-0.076}$ and $T = 0.507^{+0.063}_{-0.078}$.

concordance $\log R$ distribution. From this distribution, we can calculate an empirical CDF and compare the observed R statistic for the three pairs of observations. Nested sampling returns an error on the Bayesian evidence, which can then be propagated forward through to $\log R_{\text{obs}}$ and the tension statistics T and C . For the two in tension datasets we find $T = 2.989^{+0.167}_{-0.060}$ and $T = 2.147^{+0.056}_{-0.089}$ and for the in concordance case when both experiments observe the same signal $C = 0.864^{+0.107}_{-0.076}$ and $T = 0.507^{+0.063}_{-0.078}$. This is in agreement with our expectations, given the amplitude of the signals in the different datasets, and demonstrates that the TENSIONNET performs well. The results are summarized in Fig. 5.

B. DESI and SDSS

We next investigate the tension between the BAO cosmological constraints from the SDSS [65,66] and the recent DESI data release [67].

Before recombination when photons and baryons were coupled via Thomson scattering, oscillations were set up in the hot plasma by the competing forces of gravity and radiation pressure. Spherical density perturbations in the coupled plasma propagated outward as acoustic waves. Once the photons and the baryons decouple at recombination, these acoustic waves stop traveling through the baryon fluid and the scale of the wave is imprinted in the matter distribution. The scale of the acoustic waves at recombination is known as the sound horizon. The photons free stream and form the CMB. Since the baryons and dark

matter are coupled by gravity, the acoustic waves imprint a preferential scale for structure formation and the distance between two galaxies in the later Universe. The BAO scale is hence a standard ruler, and observations of it can be used to constrain the expansion rate of and the matter density of the Universe [27,68].

In practice, the BAO scale is estimated via the cross-correlation of the position of galaxies ξ in large surveys like SDSS and DESI and shows up as a bump in $\xi(\theta)$ and $\xi(\Delta z)$ where θ is the angular separation of galaxies and Δz the redshift separation. Angular scales on the sky θ are related to comoving physical sizes λ by

$$\theta = \frac{\lambda}{(1+z)D_A} = \frac{\lambda}{D_M}, \quad (23)$$

where D_A is the angular diameter distance and D_M is the comoving angular diameter distance also known as the transverse comoving distance. Similarly, physical size is related to redshift separation by

$$\Delta z = \frac{\lambda H(z)}{c} = \frac{\lambda}{D_H}, \quad (24)$$

where D_H is the Hubble distance, c is the speed of light and $H(z)$ is the Hubble constant as a function of redshift. From $\xi(\theta)$ and $\xi(\Delta z)$ we can approximate the angular size of the BAO at a given redshift θ_{BAO} and, given a large enough set of galaxy measurements as a function of redshift, the redshift separation Δz_{BAO} . The comoving size of the BAO

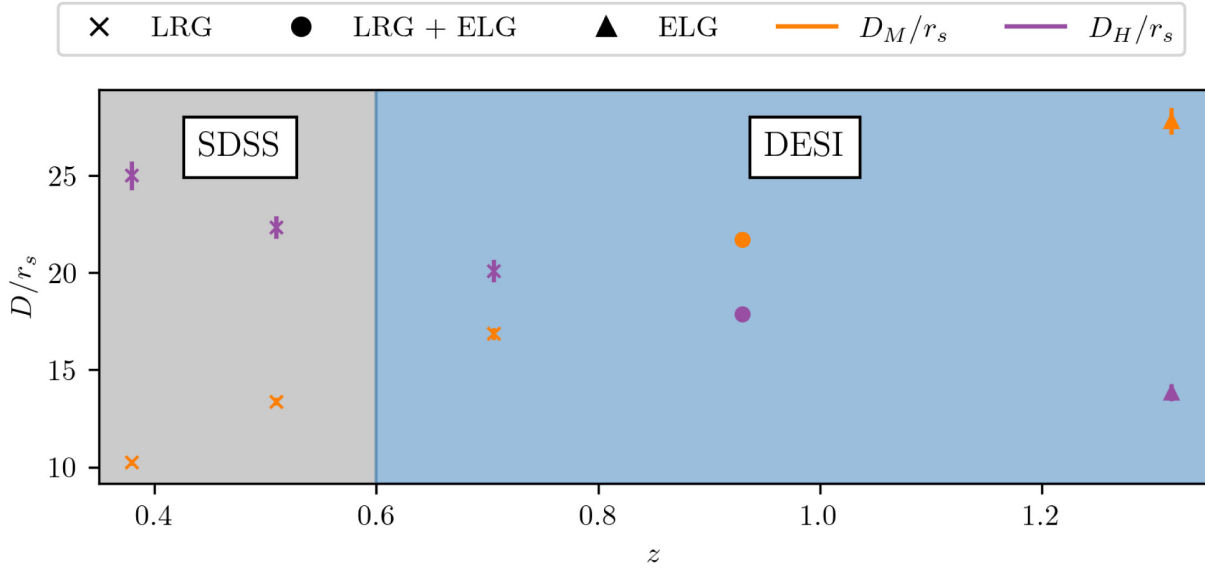


FIG. 6. Composite BAO dataset used in this work from SDSS and DESI observations. Following the discussion in [67] we curate the dataset by taking SDSS observations of the BAO scale from LRG below $z = 0.6$ (grey shaded region) and high redshift observations of luminous red galaxies and ELG from DESI (blue shaded region).

is equal to the sound horizon $\lambda = r_s$ at recombination when the photons and baryons decouple. BAO observations therefore give us a measure of D_M/r_s and D_H/r_s from which we can constrain cosmology.

The BAO signature appears in the cross-correlation of a number of different objects such as luminous red galaxies (LRG), emission line galaxies (ELG), quasars and the Lyman- α forest. Each class of objects probes a different redshift range, and measurements of D_M/r_s and D_H/r_s are ascribed to an effective redshift [68].

We generate theoretical models for the observables with CAMB [69,70], then taking advantage of the reported covariance estimates for the SDSS and DESI observations use analytic likelihoods, implemented with SciPy, to generate noisy observations of the theory model.

There is a partial overlap in the redshift range and sky coverage of SDSS and DESI, and as such the full datasets include some of the same galaxies. Therefore, the surveys are correlated and if we want to perform a joint Bayesian inference of the datasets with tools like nested sampling to recover R_{obs} we need a joint likelihood function. Although the level of correlation between the datasets has been estimated [67,71], the derivation of a joint likelihood is beyond the scope of this paper and has not yet been attempted in the literature.

An alternative approach is to build a joint SDSS and DESI dataset by selecting data points from one survey or the other at each sampled effective redshift. In [67] the authors demonstrate this idea by selecting SDSS observations below $z = 0.6$ and DESI observations above $z = 0.6$ to maximize the effective volume covered by the joint dataset. In our analysis we use

- (i) SDSS LRG at $z_{\text{eff}} = 0.38$ and 0.51 .
- (ii) DESI LRG at $z_{\text{eff}} = 0.706$.
- (iii) DESI LRG-ELG at $z_{\text{eff}} = 0.930$.
- (iv) DESI ELG at $z_{\text{eff}} = 1.317$.

and the combined dataset is shown in Fig. 6. A more complete analysis can be pursued in the future when correlated likelihoods become available. Some tension, at approximately the 3σ level, has been observed between SDSS and DESI at an effective redshift of $z_{\text{eff}} \approx 0.7$ [72], although this was not arrived at via a joint analysis but rather an assessment of the individual measurements and the correlation between the datasets. We do not expect to see this tension in our analysis, as we are just considering the DESI measurement at $z_{\text{eff}} \approx 0.7$.

We constrain the baryon density $\Omega_b h^2$, dark matter density $\Omega_c h^2$, the slope and amplitude of the matter power spectrum n_s and $\log 10^{10} A_s$ and the value of $h = \frac{H_0}{100 \text{ km s}^{-1} \text{ Mpc}^{-1}}$. We fix the value of τ to the best fit value from the Planck 2018 analysis of 0.055 [73]. The prior is uniform on $\Omega_b h^2$ between 0.01–0.085, $\Omega_c h^2$ between 0.08–0.21, n_s between 0.8–1.2, $\log 10^{10} A_s$ between 2.6–3.8 and h between 0.5–0.9. It is motivated by the prior in [15], which is somewhat motivated by the default priors for CosmoMC [74], and designed to encompass the Planck and Dark Energy Survey Y1 posteriors. As discussed in [15], however, there is nothing particularly special about this prior and in practice it could be broadened or narrowed without causing any objections in the community. For each measurement of the BAO signature D our likelihood is Gaussian, as in [27], with a covariance given by the measured covariance Σ .

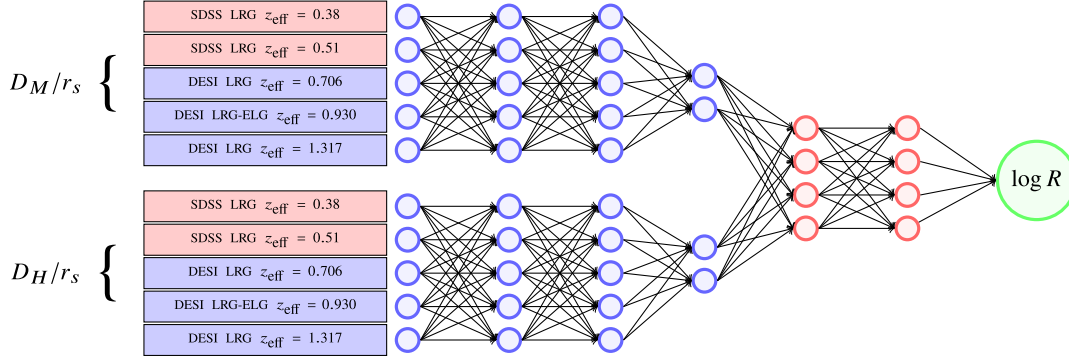


FIG. 7. Exact diagram of hidden layer structure in the DESI-SDSS TENSIONNET. We find that combining and compressing the information in the measurements of D_M/r_s and D_H/r_s from DESI and SDSS into a latent space before mixing information from the two measurements improves the performance of the NRE. The compression networks (in blue) and the NRE (in red) are trained together under the same binary cross entropy loss function.

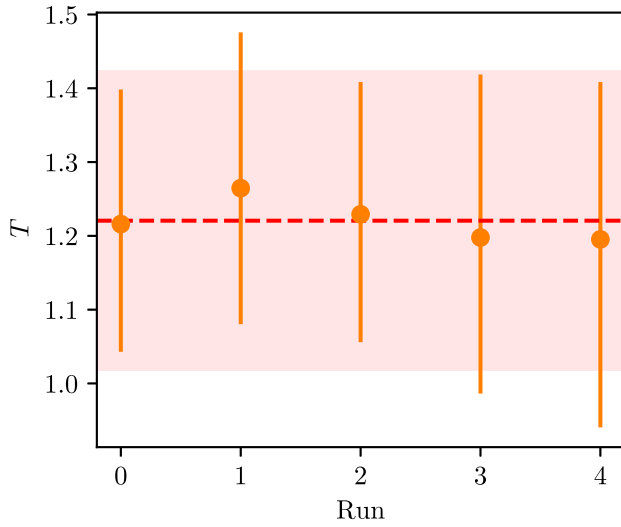


FIG. 8. We repeat training of our NRE five times on simulations of the data from SDSS and DESI and use the predicted distributions to evaluate T . If our network was poorly converged, or we had too little training data, then the recovered distribution of in concordance R values would vary significantly. As a result, the calculated value of T would be inconsistent, and we would see a large scatter in the reported values. Instead, we see that, for the SDSS + DESI analysis, T is consistent across the different training runs. On average, we find that $T = 1.22 \pm 0.20$.

The SDSS data is available at² and the DESI data has been reported in [67]. Both datasets have been collected together as part of the Cobaya cosmological likelihood code [75]. Using nested sampling and CAMB, we find $\log R_{\text{obs}} = 2.57 \pm 0.30$. Since $R_{\text{obs}} < 10$ we are not worried about the NRE saturation that was previously discussed.

To train the NRE, we generate 100,000 examples of in concordance observations from SDSS and DESI. We then

separate out 10% of these for testing and shuffle the remaining 90% to create a set of 180,000 matched and mismatched observations. These are then split into training and validation datasets of 120,600 and 59,400 (33%) observations, respectively. We use an exponentially decaying learning rate scheduler with an initial learning rate of 10^{-3} , a step size of 1000 epochs and a decay rate of 0.9. We train for a maximum of 1000 epochs with a batch size of 1000 and a patience of 50. We use L1 kernel regularization to improve the performance. We standardize the simulations at each redshift using the mean and standard deviation of the training data.

We group together the measurements of D_M/r_s from DESI and SDSS at the different effective redshifts and compress them down into a smaller latent space. We do the same with the measurements of D_H/r_s before passing them to the NRE. We find that compressing the data in this way works better than directly passing the raw data to the NRE. This initial step of keeping the measurements of D_M/r_s and D_H/r_s separate allows the NRE to learn the trends, like those seen in Fig. 6, in each variable as a function of redshift before mixing information from the two together. We found that smaller neural network architectures worked better for this problem, as there are only 10 data points in the combined set. The compression networks have three layers of 5, 5 and 2 hidden nodes and the NRE has 2 layers of 4 nodes each. The compression layers and the NRE are trained together. The architecture of the network can be seen in Fig. 7.

We train the network on the same training data five times with different random initial seeds to assess the consistency of our results. The corresponding values of T are shown in Fig. 8. We find that on average, $T = 1.22 \pm 0.20$ between the combined SDSS and DESI datasets. We show an example of the calibration performed for one of the training and calibration runs in Fig. 9 along with the constraints on Ω_m and $H_0 r_s$. We find no significant tension between the SDSS measurements of the BAO scale at $z_{\text{eff}} = 0.38$ and

²<https://www.sdss4.org/dr17/>

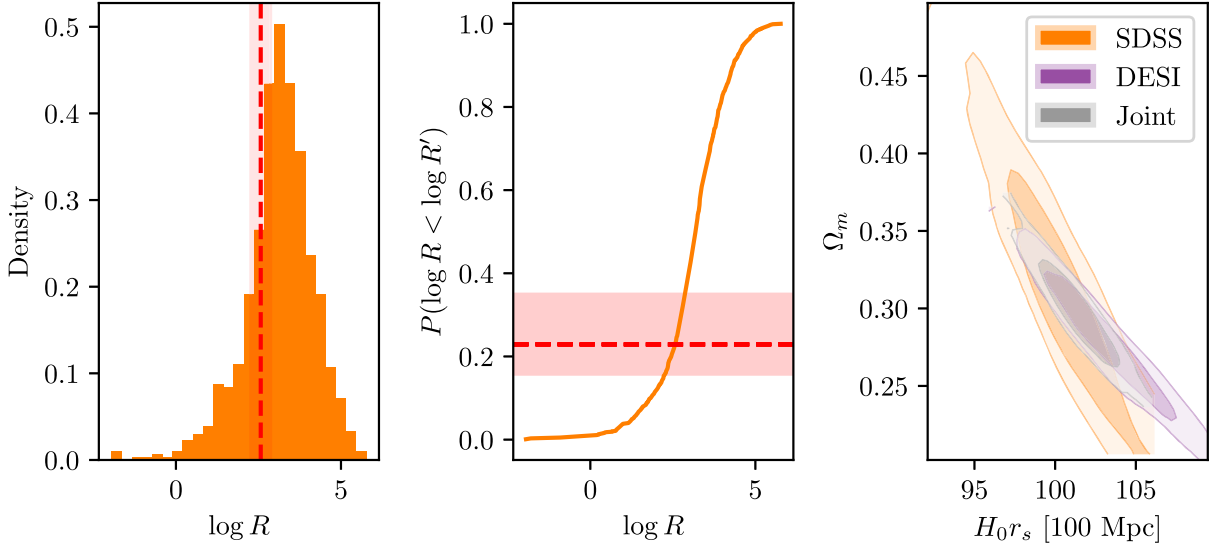


FIG. 9. Left panel: The predicted distribution of in concordance $\log R$ values for the curated SDSS and DESI dataset analyzed in this work. The red dashed line shows the value of $\log R_{\text{obs}}$ for the observed data calculated with nested sampling with an associated error. Middle panel: The CDF corresponding to the in concordance $\log R$ distribution. Calibrating $\log R_{\text{obs}}$ (red dashed line and shaded region) into σ s of tension gives $T = 1.23^{+0.21}_{-0.20}$. Right panel: The constraints on the matter overdensity Ω_m and the combination of the Hubble constant H_0 and sound horizon r_s from analysis of the DESI and SDSS datasets. Our results are slightly different to those presented in Fig. 2 of [67] because we have used a different prior and not included the quasar measurements from DESI or the Lyman- α measurements from both surveys.

0.51 and the DESI measurements at higher redshifts of $z_{\text{eff}} = 0.706, 0.930$ and 1.317 .

VIII. LIMITATIONS

As with all simulation based inference methods, the success of the TENSIONNET is dependent on how well the simulations represent the true observed datasets. In some respects, the method is also limited by the need for simulations. For example, to assess the tension between supernova observations of H_0 and CMB measurements using the R statistic and the TENSIONNET one would need to be able to simulate the observations in a consistent framework. While work is being pursued in this direction e.g. [76,77] it is a notoriously difficult problem.

It is also currently difficult to verify the output of the TENSIONNET. In practice, one could run a coverage test on the recovered distribution of $\log R$ [78]. However, this only tells you how self-consistent the recovered distribution is and not whether it is centered around the correct $\log R$ value. One way to test this is to take a number of simulated datasets in the predicted distribution and calculate their $\log R$ value via an independent method such as nested sampling. An alternative validation approach is to repeat the NRE training to check for stability as in Fig. 8.

As demonstrated in Sec. VI, the TENSIONNET is limited by the NREs ability to predict extreme values of $\log R$. Sensible choices of prior distributions can help alleviate this issue, and the validation methods discussed above can help build confidence in the predicted distribution.

IX. CONCLUSIONS

Estimating tension between different datasets is an important part of the scientific process and has become integral to the analysis of cosmological and astrophysical data. By correctly quantifying tension between different experiments, we are able to better understand our instruments and identify gaps in our knowledge. Commonly encountered examples of tension in cosmology are the H_0 and σ_8 tensions, although other examples exist, including in the field of 21-cm cosmology.

A number of ways to quantify tension have been proposed including eigentension, goodness of fit degradation and Suspiciousness and these can often be translated into σ s of tension where σ is the standard deviation of a normal distribution. A Bayesian way to quantify tension is with the tension statistic R which encodes our increased confidence in one experiment's measured data given observations from another. Formerly, R is the ratio of joint Bayesian evidence to the product of the individual evidences for two datasets under a common model and prior. It is symmetric, parametrization invariant and dimensionally consistent, however, it has a nontrivial dependence on the prior. $\log R$ is typically interpreted as indicating tension if $R \ll 1$ and concordance if $R \gg 1$ or via a Jeffery's scale, neither of which properly accounts for the prior dependence.

For any pair of experiments observing the same physics, any model for the data and any prior distribution, there is a distribution of in concordance $\log R$ values. Having access to this distribution allows you to calibrate out the prior

dependence from the observed R and robustly convert the statistic into σ s of tension or concordance. Unfortunately, for most problems, this distribution is not analytically accessible. In this paper, we have shown that it can be readily accessed with simulations of the experimental observables and neural ratio estimation.

We demonstrated the application of NREs to the calibration of R using toy examples and observations of the BAO scale from SDSS and DESI. By selecting observations of the BAO scale from each survey at specific effective redshifts, we avoid having to worry about the correlation between the observations whilst maximizing the effective volume of the combined survey. We find no significant tension between the SDSS luminous red galaxy measurements at $z_{\text{eff}} = 0.38$ and 0.51 and the DESI luminous red galaxy measurements and emission line galaxy measurements at $z_{\text{eff}} = 0.706$, 0.930 and 1.317 .

In [72] some tension has been seen between the SDSS and DESI datasets at $z_{\text{eff}} \approx 0.7$. In practice, this could be assessed with the TENSIONNET in the future should a correlated likelihood function become available for calculating the observed R with nested sampling.

Like all simulation based methods, the TENSIONNET is limited by the accuracy of the simulated observations and indeed by our ability to simulate the data in the first instance. We also find that performance of the NRE degrades as the prior widens, and sensible prior choices need to be made. We suggest that repeated training of the NRE and evaluation of R for a handful of simulations with nested sampling or an independent evidence estimation tool can be done to validate the results.

We have shown that neural ratio estimators offer a cheap and effective way to access the in concordance $\log R$ distribution needed to calibrate out the prior dependence of the R statistic. While acknowledging the limitations of this method, we believe it offers a promising step toward simulation based tension quantification. We expect that the method proposed in this paper will be broadly applicable beyond cosmology in other fields where tensions appear e.g. [79].

ACKNOWLEDGMENTS

H. T. J. B. acknowledges support from the Kavli Institute for Cosmology Cambridge and the Kavli Foundation. W. J. H. thanks the Royal Society for their support through their University Research Fellowships. T. G. J. acknowledges the support of the Science and Technology Facilities Council (UK) through Grant No. ST/V506606/1 and the Royal Society. This work used the DiRAC Data Intensive service (CSD3, Project No. ACSP289) at the University of Cambridge, managed by the University of Cambridge University Information Services on behalf of the STFC DiRAC HPC Facility [80]. The DiRAC component of CSD3 at Cambridge was funded by BEIS, UKRI and STFC capital funding and STFC operations grants. DiRAC is part of the UKRI Digital Research Infrastructure.

DATA AVAILABILITY

The code and data used in this paper are available at [81].

-
- [1] L. Knox and M. Millea, Hubble constant hunter's guide, *Phys. Rev. D* **101**, 043533 (2020).
 - [2] G. Efstathiou, A lockdown perspective on the Hubble tension (with comments from the SH0ES team), *arXiv: 2007.10716*.
 - [3] A. Amon and G. Efstathiou, A non-linear solution to the S_8 tension?, *Mon. Not. R. Astron. Soc.* **516**, 5355 (2022).
 - [4] C. Preston, A. Amon, and G. Efstathiou, A non-linear solution to the S_8 tension—II. Analysis of DES Year 3 cosmic shear, *Mon. Not. R. Astron. Soc.* **525**, 5554 (2023).
 - [5] Dark Energy Survey and Kilo-Degree Survey Collaboration, DES Y3 + KiDS-1000: Consistent cosmology combining cosmic shear surveys, *Open J. Astrophys.* **6**, 36 (2023).
 - [6] S. Singh, N. T. Jishnu, R. Subrahmanyam, N. Udaya Shankar, B. S. Girish, A. Raghunathan, R. Somashekar, K. S. Srivani, and M. Sathyanarayana Rao, On the detection of a cosmic dawn signal in the radio background, *Nat. Astron.* **6**, 607 (2022).
 - [7] R. A. Battye, T. Charnock, and A. Moss, Tension between the power spectrum of density perturbations measured on large and small scales, *Phys. Rev. D* **91**, 103508 (2015).
 - [8] W. Handley, Curvature tension: Evidence for a closed universe, *Phys. Rev. D* **103**, L041301 (2021).
 - [9] P. J. E. Peebles, Tests of cosmological models constrained by inflation, *Astrophys. J.* **284**, 439 (1984).
 - [10] L. M. Krauss and M. S. Turner, The cosmological constant is back, *Gen. Relativ. Gravit.* **27**, 1137 (1995).
 - [11] J. P. Ostriker and P. J. Steinhardt, The observational case for a low-density Universe with a non-zero cosmological constant, *Nature (London)* **377**, 600 (1995).
 - [12] G. Efstathiou, W. J. Sutherland, and S. J. Maddox, The cosmological constant and cold dark matter, *Nature (London)* **348**, 705 (1990).
 - [13] A. G. Riess *et al.*, Observational evidence from supernovae for an accelerating universe and a cosmological constant, *Astron. J.* **116**, 1009 (1998).

- [14] E. Abdalla *et al.*, Cosmology intertwined: A review of the particle physics, astrophysics, and cosmology associated with the cosmological tensions and anomalies, *J. High Energy Astrophys.* **34**, 49 (2022).
- [15] W. Handley and P. Lemos, Quantifying tensions in cosmological parameters: Interpreting the DES evidence ratio, *Phys. Rev. D* **100**, 043504 (2019).
- [16] M. Raveri, G. Zacharegkas, and W. Hu, Quantifying concordance of correlated cosmological data sets, *Phys. Rev. D* **101**, 103527 (2020).
- [17] M. Raveri and C. Doux, Non-Gaussian estimates of tensions in cosmological parameters, *Phys. Rev. D* **104**, 043504 (2021).
- [18] M. Raveri and W. Hu, Concordance and discordance in cosmology, *Phys. Rev. D* **99**, 043506 (2019).
- [19] Y. Park and E. Rozo, Concordance cosmology?, *Mon. Not. R. Astron. Soc.* **499**, 4638 (2020).
- [20] T. Charnock, R. A. Battye, and A. Moss, Planck data versus large scale structure: Methods to quantify discordance, *Phys. Rev. D* **95**, 123535 (2017).
- [21] DES Collaboration, Assessing tension metrics with dark energy survey and Planck data, *Mon. Not. R. Astron. Soc.* **505**, 6179 (2021).
- [22] E. Saraivanov, K. Zhong, V. Miranda, S. S. Boruah, T. Eifler, and E. Krause, Attention-based neural network emulators for multi-probe data vectors Part II: Assessing tension metrics, *Phys. Rev. D* **111**, 123520 (2025).
- [23] P. Marshall, N. Rajguru, and A. Slosar, Bayesian evidence as a tool for comparing datasets, *Phys. Rev. D* **73**, 067302 (2006).
- [24] R. Trotta, Bayes in the sky: Bayesian inference and model selection in cosmology, *Contemp. Phys.* **49**, 71 (2008).
- [25] S. Seehars, S. Grandis, A. Amara, and A. Refregier, Quantifying concordance in cosmology, *Phys. Rev. D* **93**, 103507 (2016).
- [26] M. Cortês and A. R. Liddle, On data set tensions and signatures of new cosmological physics, *Mon. Not. R. Astron. Soc.* **531**, L52 (2024).
- [27] A. Cuceu, J. Farr, P. Lemos, and A. Font-Ribera, Baryon Acoustic Oscillations and the Hubble constant: Past, present and future, *J. Cosmol. Astropart. Phys.* **10** (2019) 044.
- [28] K. Cranmer, J. Brehmer, and G. Louppe, The frontier of simulation-based inference, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30055 (2020).
- [29] B. Miller, A. Cole, P. Forré, G. Louppe, and C. Weniger, Truncated marginal neural ratio estimation, *Adv. Neural Inf. Process. Syst.* **34**, 129 (2021), <https://proceedings.neurips.cc/paper/2021/hash/01632f7b7a127233fa1188bd6c2e42e1-Abstract.html>.
- [30] A. Cole, B. K. Miller, S. J. Witte, M. X. Cai, M. W. Grootes, F. Nattino, and C. Weniger, Fast and credible likelihood-free cosmology with truncated marginal neural ratio estimation, *J. Cosmol. Astropart. Phys.* **09** (2022) 004.
- [31] J. Skilling, Nested sampling for general Bayesian computation, *Bayesian Anal.* **1**, 833 (2006).
- [32] G. Ashtyn *et al.*, Nested sampling for physical scientists, *Nat. Rev. Methods Primers* **2**, 39 (2022).
- [33] R. Trotta, Applications of Bayesian model selection to cosmological parameters, *Mon. Not. R. Astron. Soc.* **378**, 72 (2007).
- [34] A. Heavens, Y. Fantaye, A. Mootoovaloo, H. Eggers, Z. Hosenie, S. Kroon, and E. Sellentin, Marginal likelihoods from Monte Carlo Markov chains, [arXiv:1704.03472](https://arxiv.org/abs/1704.03472).
- [35] R. Srinivasan, M. Crisostomi, R. Trotta, E. Barausse, and M. Breschi, floZ: Evidence estimation from posterior samples with normalizing flows, *Phys. Rev. D* **110**, 123007 (2024).
- [36] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, emcee: The MCMC Hammer, *Publ. Astron. Soc. Pac.* **125**, 306 (2013).
- [37] A. Polanska, M. A. Price, A. Spurio Mancini, and J. D. McEwen, Learned harmonic mean estimation of the marginal likelihood with normalizing flows, [arXiv:2307.00048](https://arxiv.org/abs/2307.00048).
- [38] Dark Energy Survey Collaboration, Dark Energy Survey year 1 results: Cosmological constraints from galaxy clustering and weak lensing, *Phys. Rev. D* **98**, 043526 (2018).
- [39] H. Jeffreys, *Theory of Probability*, International Series of Monographs on Physics (Clarendon Press, New York, 1983).
- [40] <https://github.com/handley-lab/lspi>.
- [41] S. R. Furlanetto, S. P. Oh, and F. H. Briggs, Cosmology at low frequencies: The 21 cm transition and the high-redshift Universe, *Phys. Rep.* **433**, 181 (2006).
- [42] R. Barkana, The rise of the first stars: Supersonic streaming, radiative feedback, and 21-cm cosmology, *Phys. Rep.* **645**, 1 (2016).
- [43] *The Cosmic 21-cm Revolution*, edited by A. Mesinger (IOP Publishing, 2019), pp. 2514–3433.
- [44] A. Liu and J. R. Shaw, Data analysis for precision 21 cm cosmology, *Publ. Astron. Soc. Pac.* **132**, 062001 (2020).
- [45] J. D. Bowman, A. E. E. Rogers, R. A. Monsalve, T. J. Mozdzen, and N. Mahesh, An absorption profile centred at 78 megahertz in the sky-averaged spectrum, *Nature (London)* **555**, 67 (2018).
- [46] S. Singh, N. T. Jishnu, R. Subrahmanyam, N. Udaya Shankar, B. S. Girish, A. Raghunathan, R. Somashekar, K. S. Srivani, and M. Sathyanarayana Rao, On the detection of a cosmic dawn signal in the radio background, *Nat. Astron.* **6**, 607 (2022).
- [47] E. de Lera Acedo *et al.*, The REACH radiometer for detecting the 21-cm hydrogen signal from redshift $z \approx 7.5 - 28$, *Nat. Astron.* **6**, 984 (2022).
- [48] J. Mirocha, Decoding the x-ray properties of pre-reionization era sources, *Mon. Not. R. Astron. Soc.* **443**, 1211 (2014).
- [49] A. Mesinger, S. Furlanetto, and R. Cen, 21 cmfast: A fast, seminumerical simulation of the high-redshift 21-cm signal, *Mon. Not. R. Astron. Soc.* **411**, 955 (2011).
- [50] I. Reis, A. Fialkov, and R. Barkana, High-redshift radio galaxies: A potential new source of 21-cm fluctuations, *Mon. Not. R. Astron. Soc.* **499**, 5993 (2020).
- [51] I. Reis, A. Fialkov, and R. Barkana, The subtlety of Ly α photons: Changing the expected range of the 21-cm signal, *Mon. Not. R. Astron. Soc.* **506**, 5479 (2021).
- [52] T. Gessey-Jones, N. S. Sartorio, A. Fialkov, G. M. Mirouh, M. Magg, R. G. Izzard, E. de Lera Acedo, W. J. Handley, and R. Barkana, Impact of the primordial stellar initial mass function on the 21-cm signal, *Mon. Not. R. Astron. Soc.* **516**, 841 (2022).

- [53] S. Sikder, R. Barkana, A. Fialkov, and I. Reis, Strong 21-cm fluctuations and anisotropy due to the line-of-sight effect of radio galaxies at cosmic dawn, *Mon. Not. R. Astron. Soc.* **527**, 10975 (2024).
- [54] S. Pochinda, T. Gessey-Jones, H. T. J. Bevins, A. Fialkov, S. Heimersheim, I. Abril-Cabezas, E. de Lera Acedo, S. Singh, S. Sikder, and R. Barkana, Constraining the properties of Population III galaxies with multiwavelength observations, *Mon. Not. R. Astron. Soc.* **531**, 1113 (2024).
- [55] T. Gessey-Jones, S. Pochinda, H. T. J. Bevins, A. Fialkov, W. J. Handley, E. de Lera Acedo, S. Singh, and R. Barkana, On the constraints on superconducting cosmic strings from 21-cm cosmology, *Mon. Not. R. Astron. Soc.* **524**, 519 (2024).
- [56] J. B. Muñoz, An effective model for the cosmic-dawn 21-cm signal, *Mon. Not. R. Astron. Soc.* **523**, 2587 (2023).
- [57] S. Mittal, ECHO21: Exploring the cosmos with hydrogen observation, [arXiv:2503.11762](https://arxiv.org/abs/2503.11762).
- [58] D. Anstey, E. de Lera Acedo, and W. Handley, A general Bayesian framework for foreground modelling and chromaticity correction for global 21 cm experiments, *Mon. Not. R. Astron. Soc.* **506**, 2041 (2021).
- [59] H. T. J. Bevins, W. J. Handley, A. Fialkov, E. de Lera Acedo, L. J. Greenhill, and D. C. Price, MAXSMOOTH: Rapid maximally smooth function fitting with applications in global 21-cm cosmology, *Mon. Not. R. Astron. Soc.* **502**, 4405 (2021).
- [60] R. Hills, G. Kulkarni, P. D. Meerburg, and E. Puchwein, Concerns about modelling of the EDGES data, *Nature (London)* **564**, E32 (2018).
- [61] S. Singh and R. Subrahmanyam, The redshifted 21 cm signal in the EDGES low-band spectrum, *Astrophys. J.* **880**, 26 (2019).
- [62] P. H. Sims and J. C. Pober, Testing for calibration systematics in the EDGES low-band data using Bayesian model selection, *Mon. Not. R. Astron. Soc.* **492**, 22 (2020).
- [63] W. J. Handley, M. P. Hobson, and A. N. Lasenby, POLYCHORD: Next-generation nested sampling, *Mon. Not. R. Astron. Soc.* **453**, 4384 (2015).
- [64] W. J. Handley, M. P. Hobson, and A. N. Lasenby, POLYCHORD: Nested sampling for cosmology, *Mon. Not. R. Astron. Soc.* **450**, L61 (2015).
- [65] S. Alam *et al.*, The eleventh and twelfth data releases of the Sloan Digital Sky Survey: Final data from SDSS-III, *Astrophys. J. Suppl. Ser.* **219**, 12 (2015).
- [66] R. Ahumada *et al.*, The 16th data release of the Sloan Digital Sky Surveys: First release from the APOGEE-2 southern survey and full release of eBOSS spectra, *Astrophys. J. Suppl. Ser.* **249**, 3 (2020).
- [67] A. Adame, J. Aguilar, S. Ahlen, S. Alam, D. Alexander, M. Alvarez, O. Alves, A. Anand, U. Andrade, E. Armengaud *et al.*, Desi 2024 VI: Cosmological constraints from the measurements of baryon acoustic oscillations, *J. Cosmol. Astropart. Phys.* **02** (2025) 021.
- [68] B. Bassett and R. Hlozek, in *Dark Energy: Observational and Theoretical Approaches*, edited by P. Ruiz-Lapuente (Cambridge University Press, Cambridge, UK, 2010), p. 246; B. Bassett and R. Hlozek, [arXiv:0910.5224](https://arxiv.org/abs/0910.5224).
- [69] A. Lewis, A. Challinor, and A. Lasenby, Efficient computation of CMB anisotropies in closed FRW models, *Astrophys. J.* **538**, 473 (2000).
- [70] A. Lewis and S. Bridle, Cosmological parameters from CMB and other data: A Monte Carlo approach, *Phys. Rev. D* **66**, 103511 (2002).
- [71] DESI Collaboration, DESI 2024 IV: Baryon acoustic oscillations from the Lyman Alpha Forest, *J. Cosmol. Astropart. Phys.* **01** (2025) 124.
- [72] DESI Collaboration, DESI 2024 III: Baryon acoustic oscillations from galaxies and quasars, *J. Cosmol. Astropart. Phys.* **04** (2025) 012.
- [73] Planck Collaboration, Planck 2018 results. VI. Cosmological parameters, *Astron. Astrophys.* **641**, A6 (2020).
- [74] <https://cosmologist.info/cosmomc/>.
- [75] https://cobaya.readthedocs.io/en/latest/likelihood_bao.html.
- [76] D. J. Watts *et al.*, COSMOGLOBE DR1 results. I. Improved Wilkinson Microwave Anisotropy Probe maps through Bayesian end-to-end analysis, *Astron. Astrophys.* **679**, A143 (2023).
- [77] K. Karchev, M. Grayling, B. M. Boyd, R. Trotta, K. S. Mandel, and C. Weniger, SIDE-real: Supernova Ia dust extinction with truncated marginal neural ratio estimation applied to real data, *Mon. Not. R. Astron. Soc.* **530**, 3881 (2024).
- [78] P. Lemos, A. Coogan, Y. Hezaveh, and L. Perreault-Levasseur, in *International Conference on Machine Learning* (PMLR, 2023), pp. 19256–19273.
- [79] CDF Collaboration, High-precision measurement of the W boson mass with the CDF II detector, *Science* **376**, 170 (2022).
- [80] www.dirac.ac.uk
- [81] <https://github.com/htjb/tension-networks>