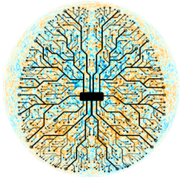
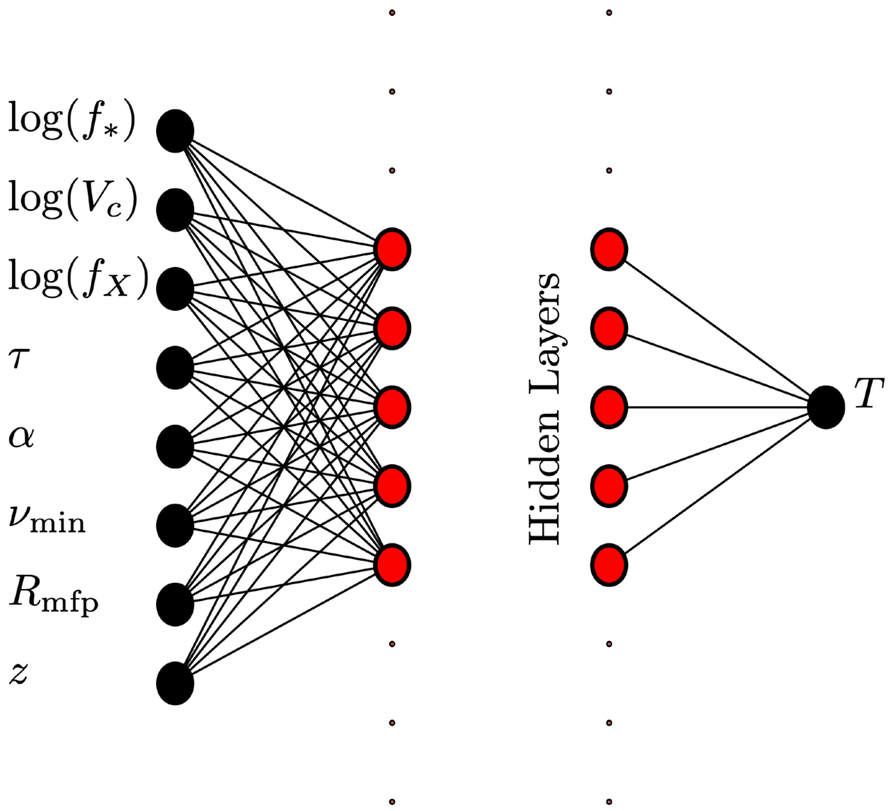
On the accuracy of posterior recovery with neural network emulators

Harry Bevins
With Thomas Gessey-Jones and Will Handley


Contents

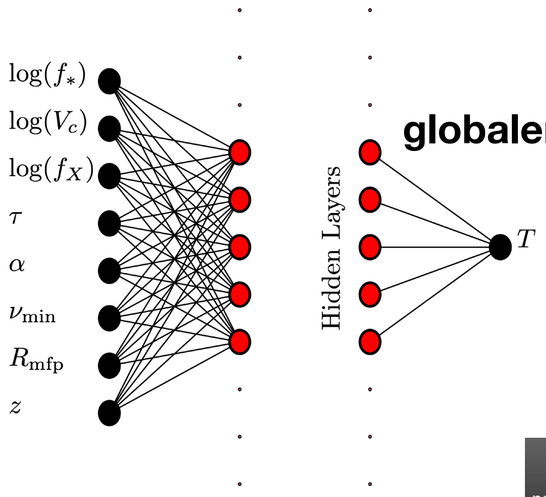
- Motivation
- Emulator Accuracy
- Experiments
- Comparison with previous work



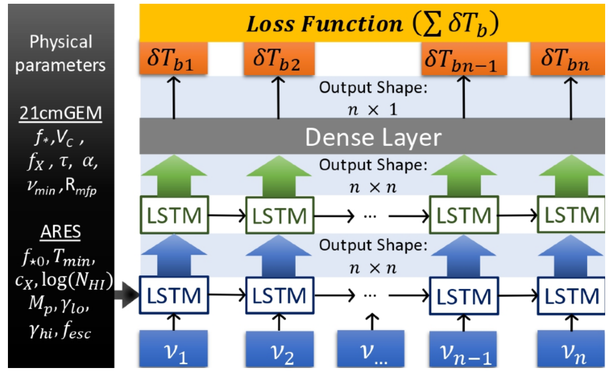
Motivation

Emulators in 21cm Cosmology

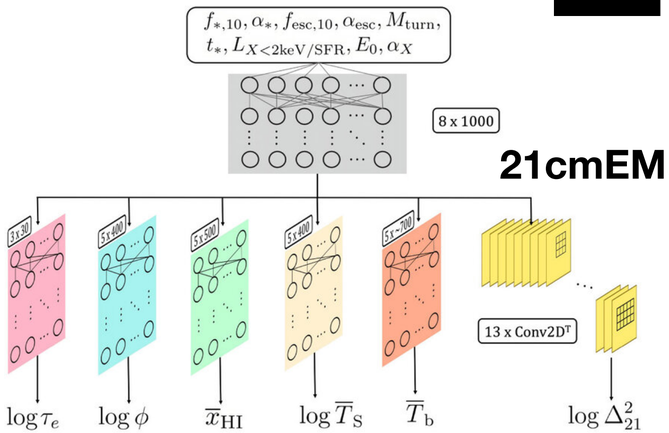
- Neural network emulators are really important in 21cm Cosmology
- If we want to perform inference on our data then we need fast likelihood functions
- Semi-numerical simulations of the signal take order hours to run per parameter set
- Idea is to train emulators on example simulations and use these in the likelihood functions



globalemu [Bevins+21]



21cmLSTM [Dorigo Jones+2024]



21cmEMU [Breitman+ 2023]

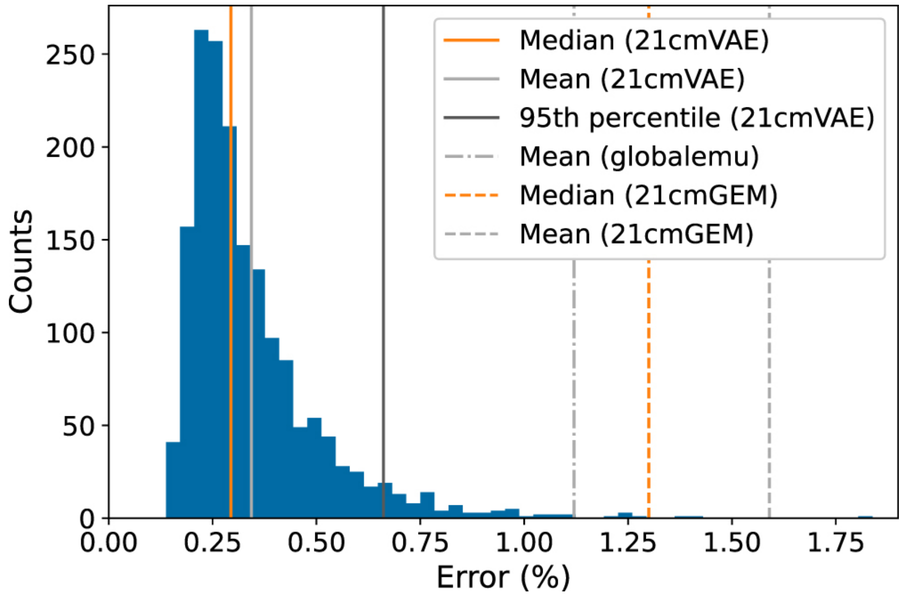
Also 21cmGEM [Cohen+2017] And 21cmVAE [Bye+2022]

Defining required accuracy

- We measure accuracy by evaluating the networks on a test data set
- Typically we do this with something like RMSE

$$\epsilon = \sqrt{\frac{1}{N_\nu} \sum_i^{N_\nu} (T_{\rm true}(\nu) - T_{\rm pred}(\nu))^2}$$

- But what average value of ϵ over the test data is good enough?
- Generally the field has worked with “rules of thumb”
- e.g. globalemu paper suggested $\bar{\epsilon} \approx 0.1\sigma$



Impact on posterior recovery?

- What we are really interested in is how well can we recover the posteriors if we use an emulator rather than the full simulation?
- Is $\bar{\epsilon} \approx 0.1\sigma$ good enough?
- Dorigo Jones+23 tried to answer this by comparing posteriors recovered with globalemu and simulations with ARES
- ARES is a 1D radiative transfer code
- Evaluates in about 1s

OPEN ACCESS

Validating Posteriors Obtained by an Emulator When Jointly Fitting Mock Data of the Global 21 cm Signal and High-*z* Galaxy UV Luminosity Function

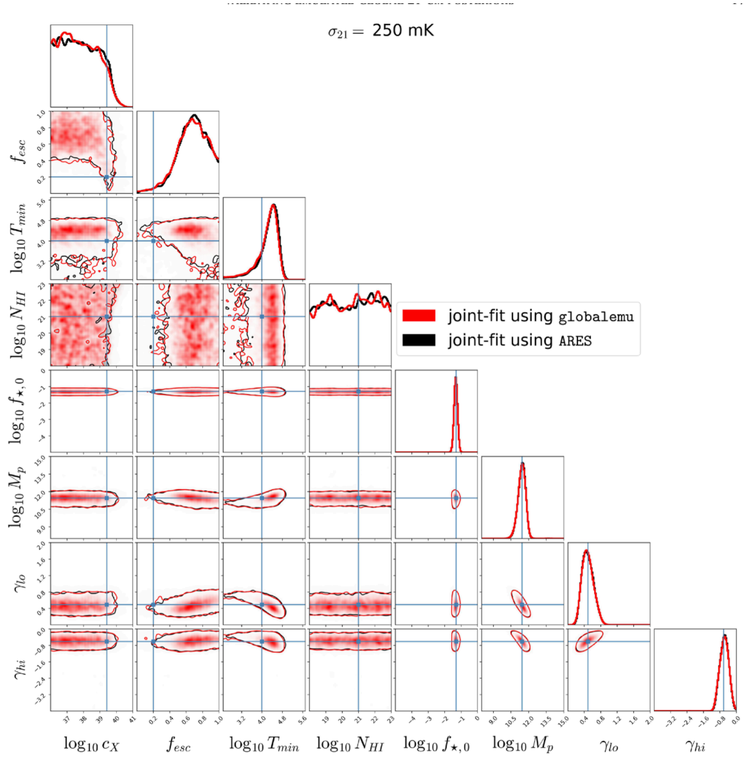
J. Dorigo Jones[1], D. Rapetti[1,2,3], J. Mirocha[4,5], J. J. Hibbard[1], J. O. Burns[1], and N. Bassett[1]

Published 2023 December 5 • © 2023. The Author(s). Published by the American Astronomical Society.

The Astrophysical Journal, Volume 959, Number 1

Citation J. Dorigo Jones *et al* 2023 *ApJ* **959** 49

DOI 10.3847/1538-4357/ad003e



Dorigo Jones+23

- Generated a fiducial ARES signal and trained a version of globalemu on around 24,000 simulations
- Tested the accuracy on around 2,000 simulations
- Then added different levels of noise to the signal with standard deviations of 5, 10, 25, 50 and 250 mK
- Performed inference on these mock data sets directly with ARES and with the emulator
- Compared the recovered posteriors
- Work complicated by inclusion of constraints from UVLF

OPEN ACCESS

Validating Posteriors Obtained by an Emulator When Jointly Fitting Mock Data of the Global 21 cm Signal and High-*z* Galaxy UV Luminosity Function

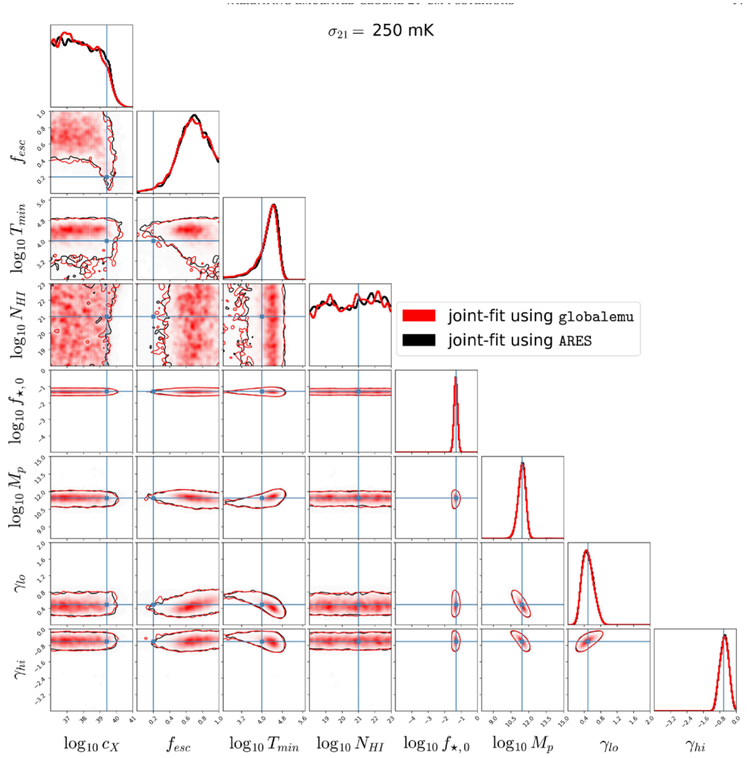
J. Dorigo Jones[1], D. Rapetti[1,2,3], J. Mirocha[4,5], J. J. Hibbard[1], J. O. Burns[1], and N. Bassett[1]

Published 2023 December 5 • © 2023. The Author(s). Published by the American Astronomical Society.

The Astrophysical Journal, Volume 959, Number 1

Citation J. Dorigo Jones *et al* 2023 *ApJ* **959** 49

DOI 10.3847/1538-4357/ad003e



Dorigo Jones+23

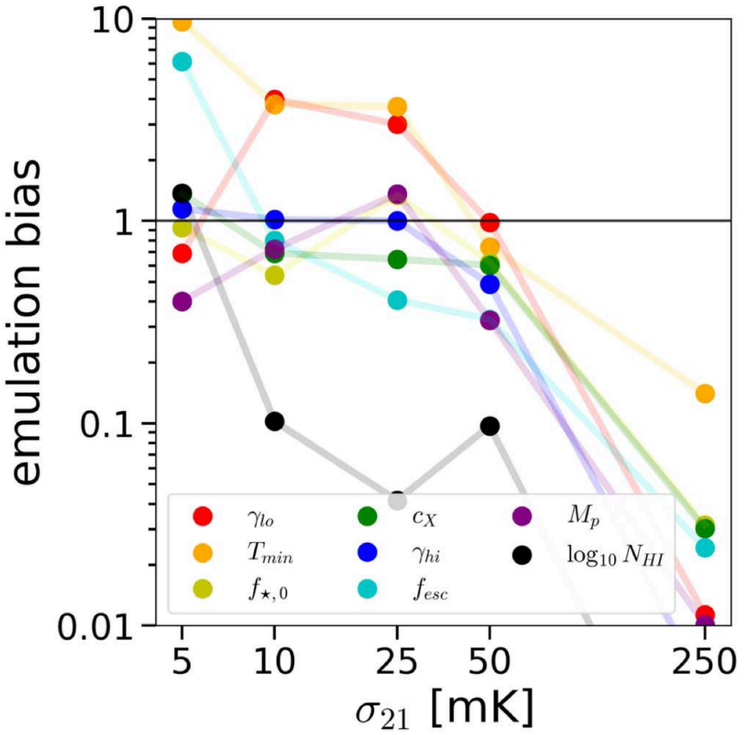
- Measured posterior accuracy with two metrics

$$\text{emulator bias} = \frac{|\mu_{\text{globalemu}} - \mu_{\text{ARES}}|}{\sigma_{\text{ARES}}}$$

$$\text{true bias} = \frac{|\mu_{\text{ARES}} - \theta_0|}{\sigma_{\text{ARES}}}$$

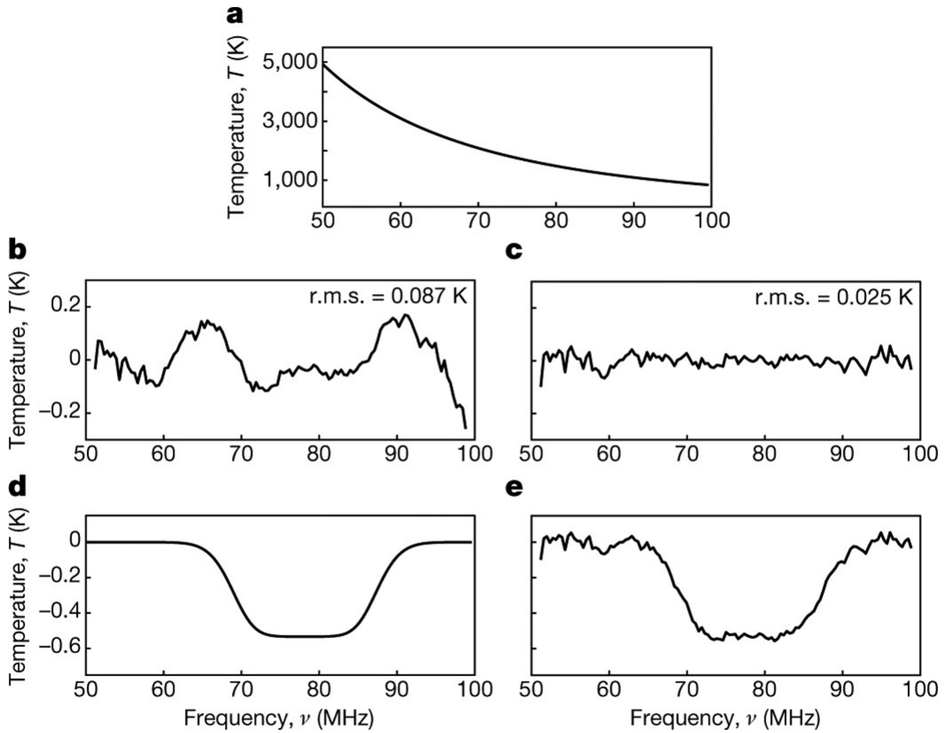
- Not really interested in true bias

- But they concluded that even for $\bar{\epsilon} \approx 0.05\sigma$ they can't accurately recover the posteriors with globalemu



Why this is concerning?

- Most experimentalists agree that we need to go down to around 25 mK noise to confidently detect the 21cm signal
- Noise level in the EDGES data is around 20 mK
- Most emulators have $\bar{\epsilon} \approx 1\ \mathrm{mK} \approx 0.05 \times 25\mathrm{mK}$ and it seems challenging to go beyond this
- We need emulators to fit physical signals to data sets

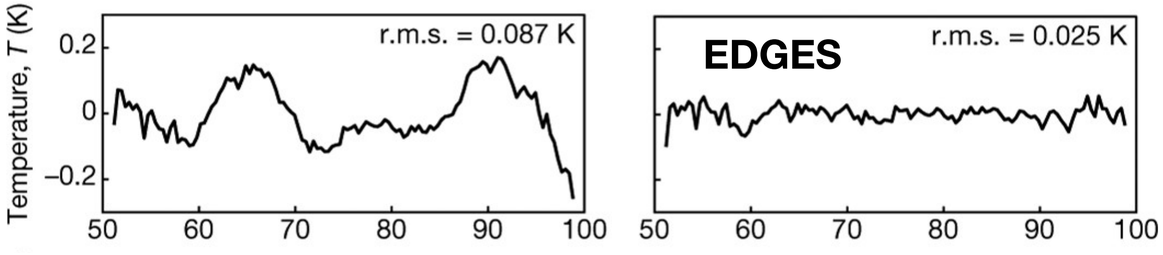


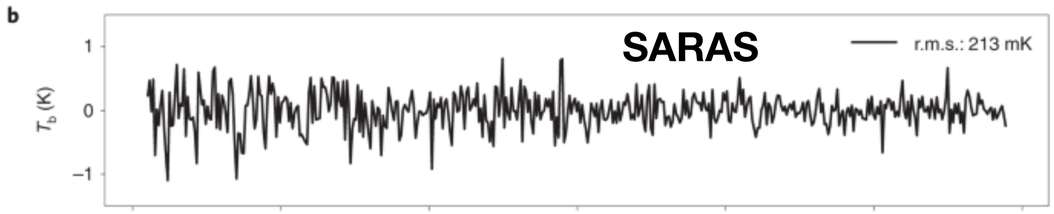
Why this seems wrong?

- We mostly assume a Gaussian likelihood function (really should be radiometric but we are not there as a field yet)
- For a Gaussian likelihood we can say

$$\sigma^2 = \sigma^2_{\text{instrument}} + \bar{\epsilon}^2$$

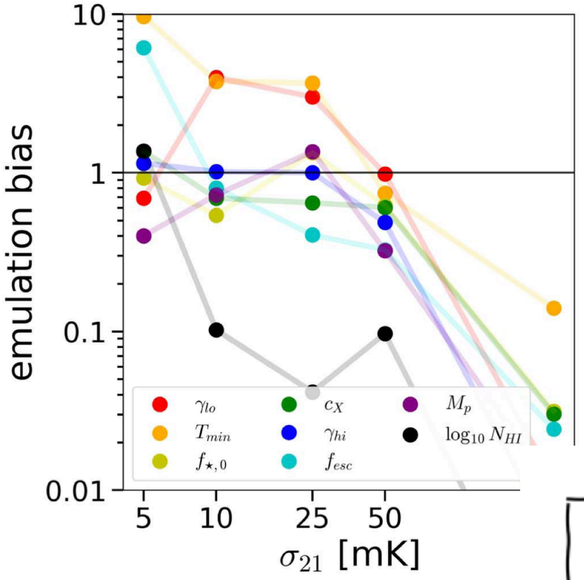
- To account for uncertainty in the emulator
- But for typical values of $\sigma_{\text{instrument}} \approx 25$ mK and $\bar{\epsilon} = 1$ mK then the emulator error is massively subdominant
- So we wouldn't expect it to impact the posterior significantly…
- Admittedly this assumes uncorrelated errors

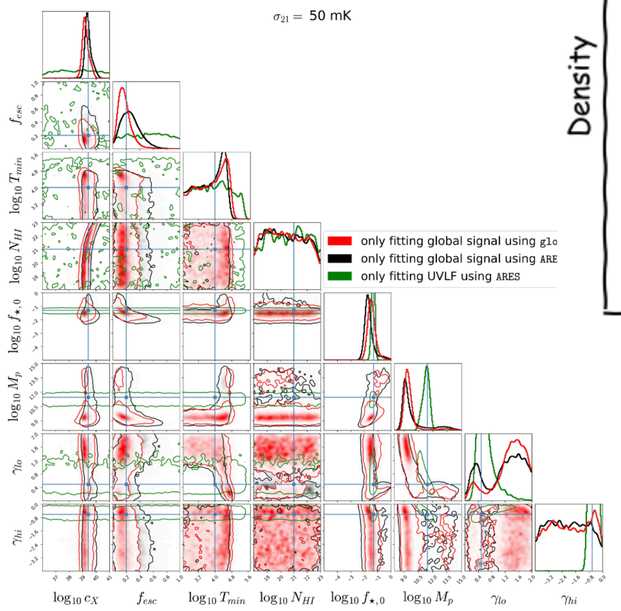


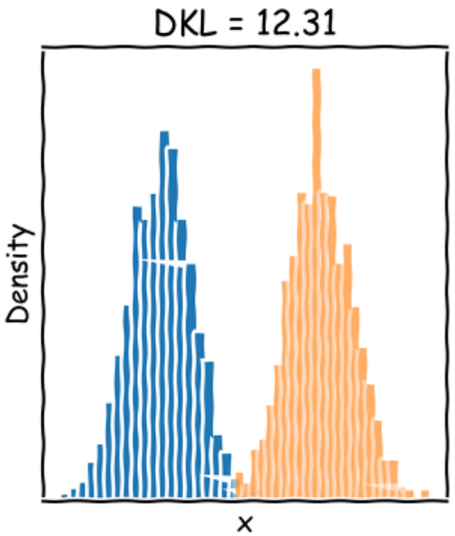


What we wanted to do?

- We wanted to repeat some of the analysis that Dorigo Jones+23 did
- And see if we could come up with a better metric to determine the posterior bias incurred from using emulators
- Address concerns that they raised about the globalemu preprocessing steps
- Do this without the UV luminosity function so that the comparison is clearer
- Make our code public!







Emulator Accuracy

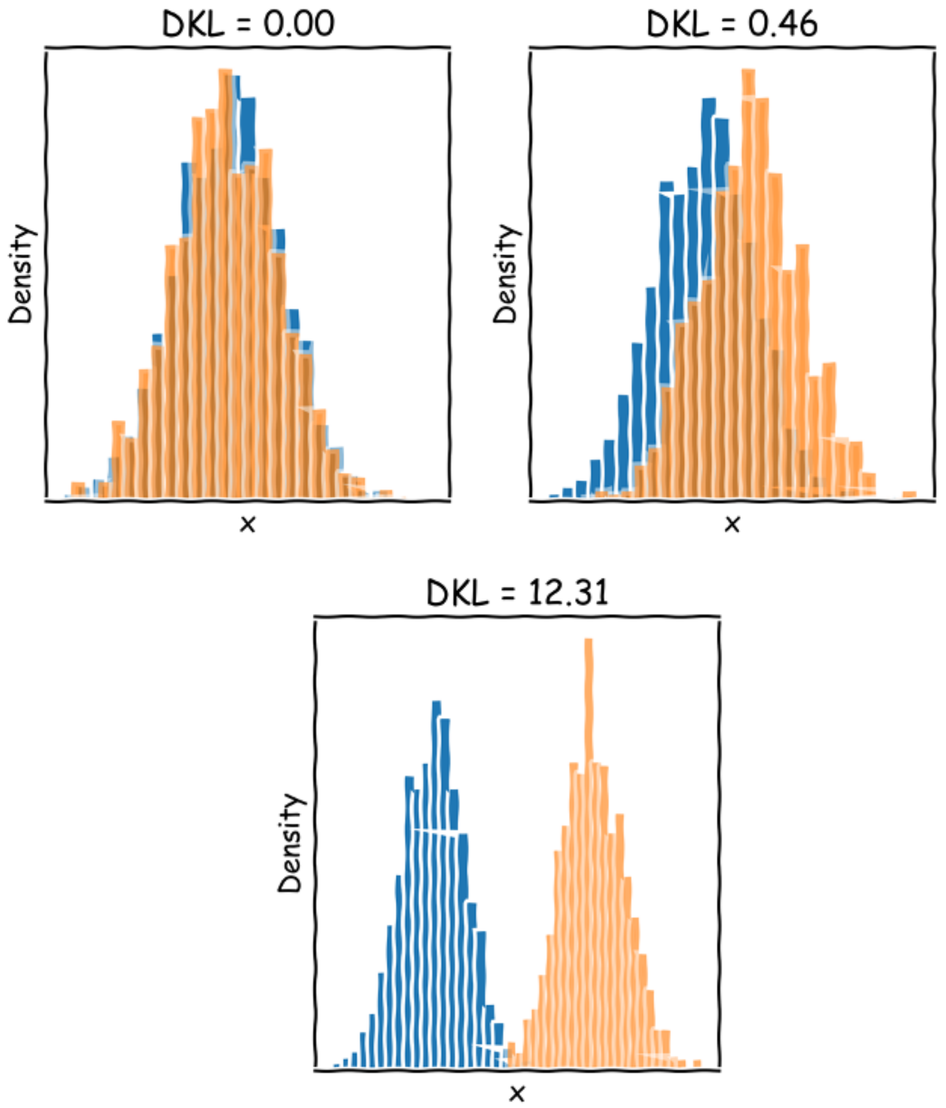
Measuring the impact of the emulator

- The emulator bias defined in Dorigo Jones+23 is fine but its only really considers the difference in 1D

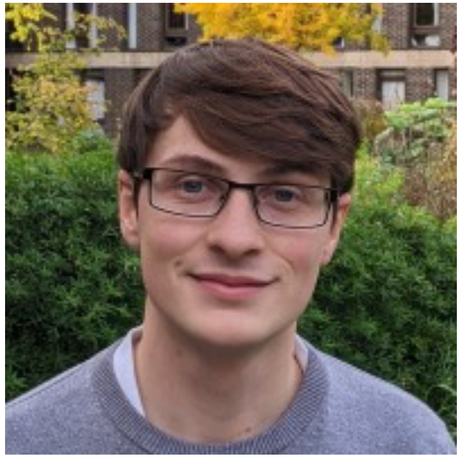
- A more comprehensive measure of the difference between the true and emulated posteriors is the Kullback-Leibler Divergence

$$D_{\mathrm{KL}} = \int P \log\left(\frac{P}{P_\epsilon}\right) d\theta$$

- The issue is that we typically do not have access to P else we wouldn't be interested in emulators



Measuring the impact of the emulator



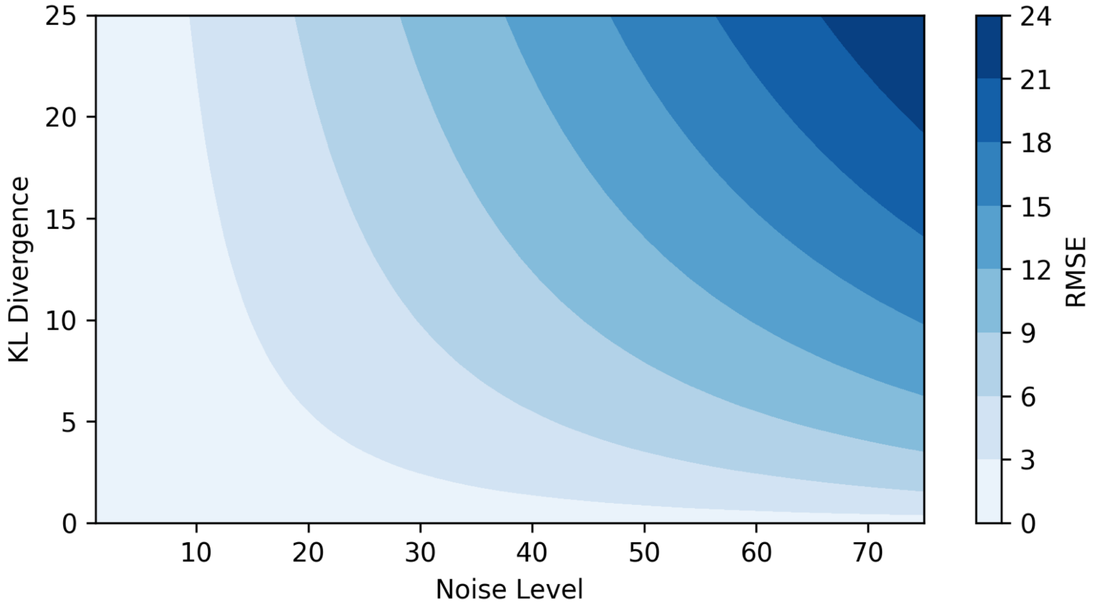
- Can make progress if we assume:
  - A linear model and linear emulator error

$$M(\theta) = M\theta + m \text{ and } E(\theta) = E\theta + \epsilon$$
$$\text{Such that } M_\epsilon(\theta) = (M + E)\theta + (m + \epsilon)$$

  - A gaussian likelihood
  - A uniform prior
  - White noise and $E \ll M$

- Can do lots of maths (I can share this with people if they are interested) and you end up with

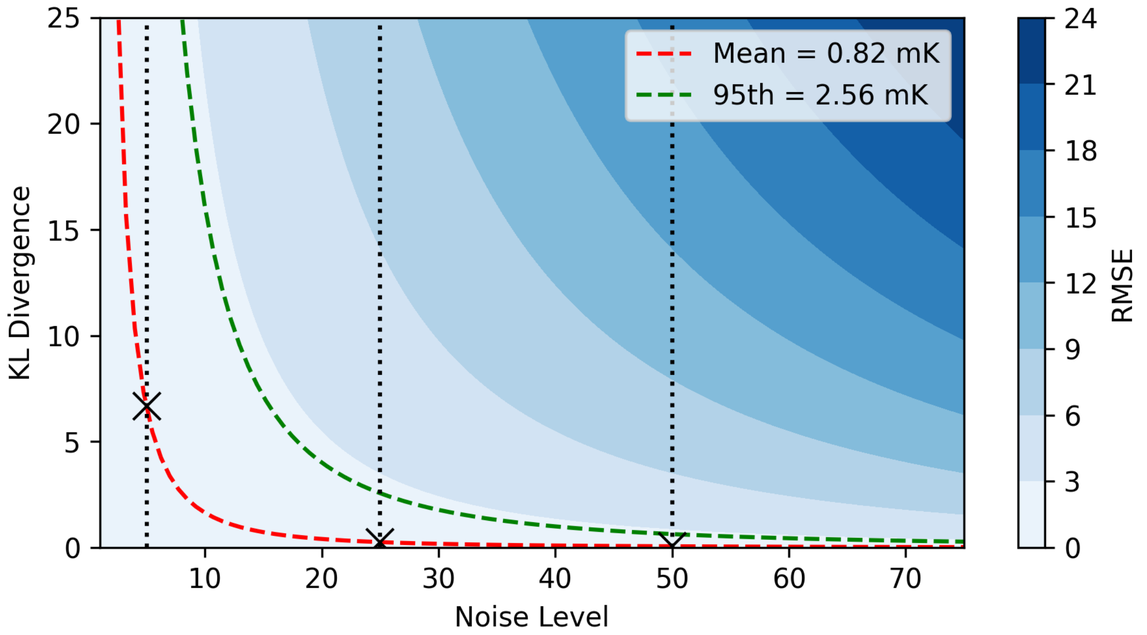
$$D_{\mathrm{KL}}(P\,||\,P_\epsilon) \leq \frac{N_d}{2}\left(\frac{\mathrm{RMSE}}{\sigma}\right)^2$$



Measuring the impact of the emulator

$$D_{\mathrm{KL}}(P\,||\,P_\epsilon) \leq \frac{N_d}{2}\left(\frac{\mathrm{RMSE}}{\sigma}\right)^2$$

- An approximate upper bound on the KL divergence between P and P_ϵ given an emulator error RMSE, the noise in the data σ and the number of data points N_d
- Predictive function that can be used both to justify but also predict the required accuracy of an emulator

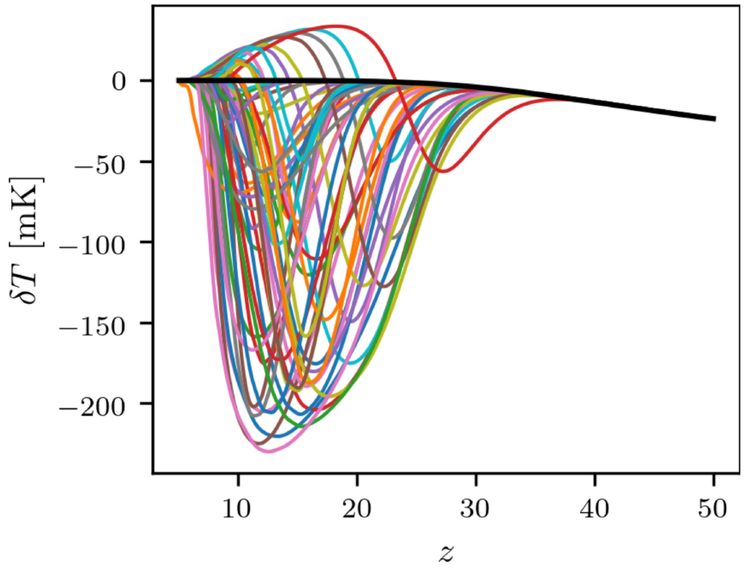


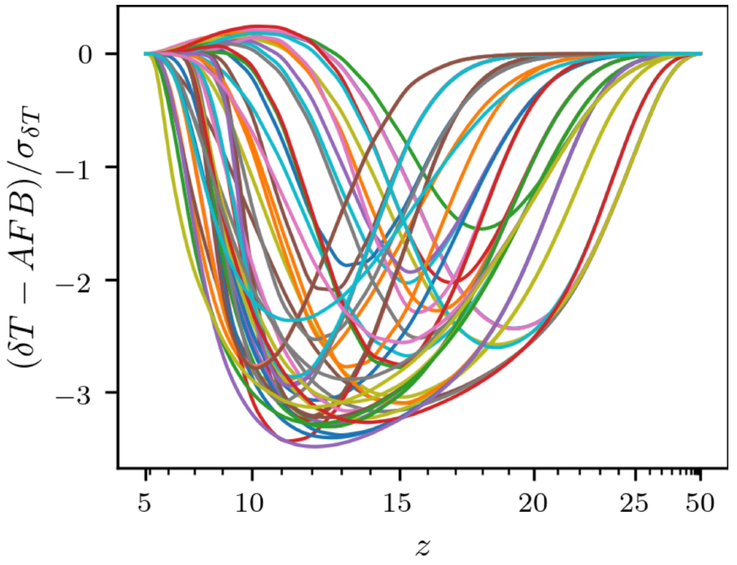
Experiments

globalemu Preprocessing

- In Dorigo Jones+23 they claimed that the preprocessing hindered the performance of globalemu
- Astrophysics free baseline subtraction - removing the common astrophysics free part of the 21cm signal from the training data
- Resampling - increasing the redshift/frequency sampling in the region where the training data varies the most
- Dorigo Jones+ shared the training data with us

**They did use an older version of globalemu

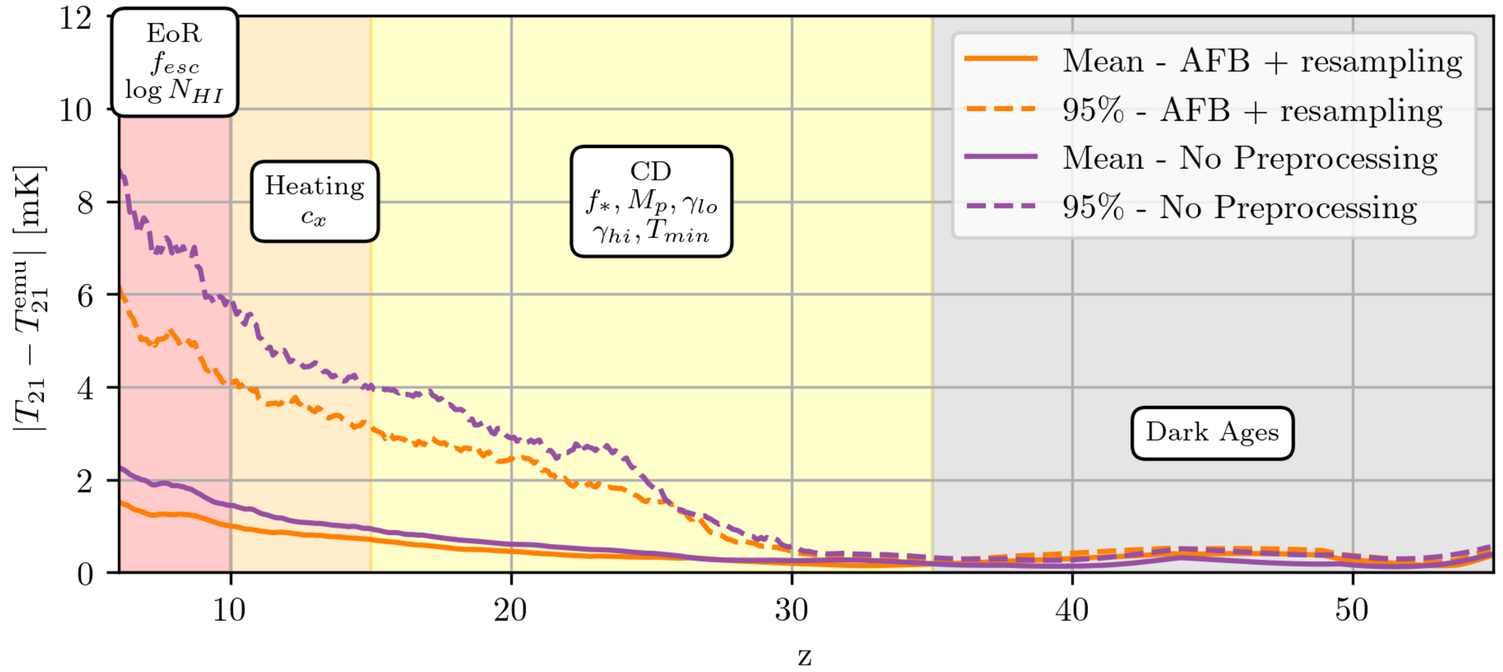




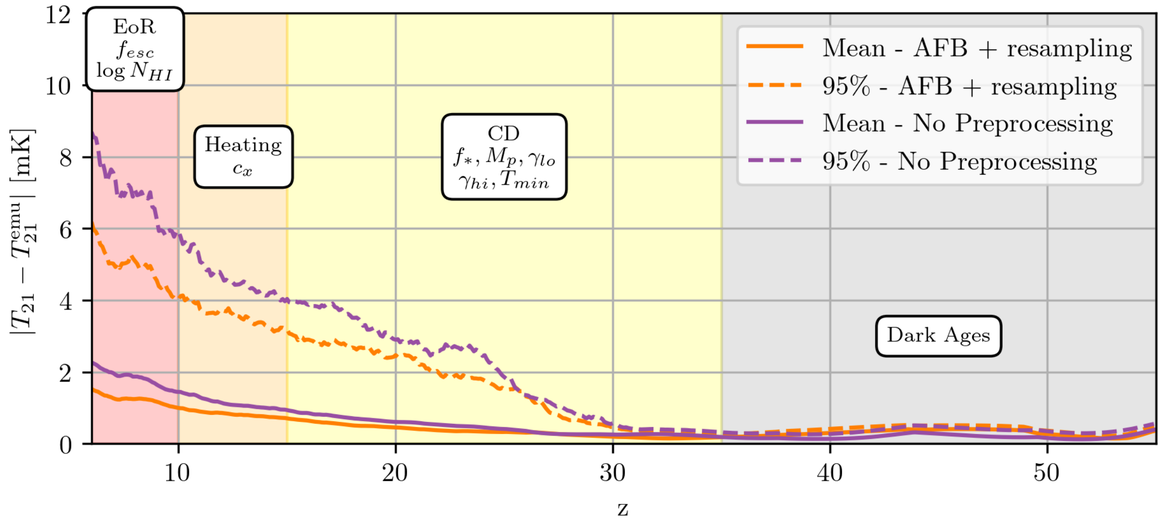
globalemu Preprocessing

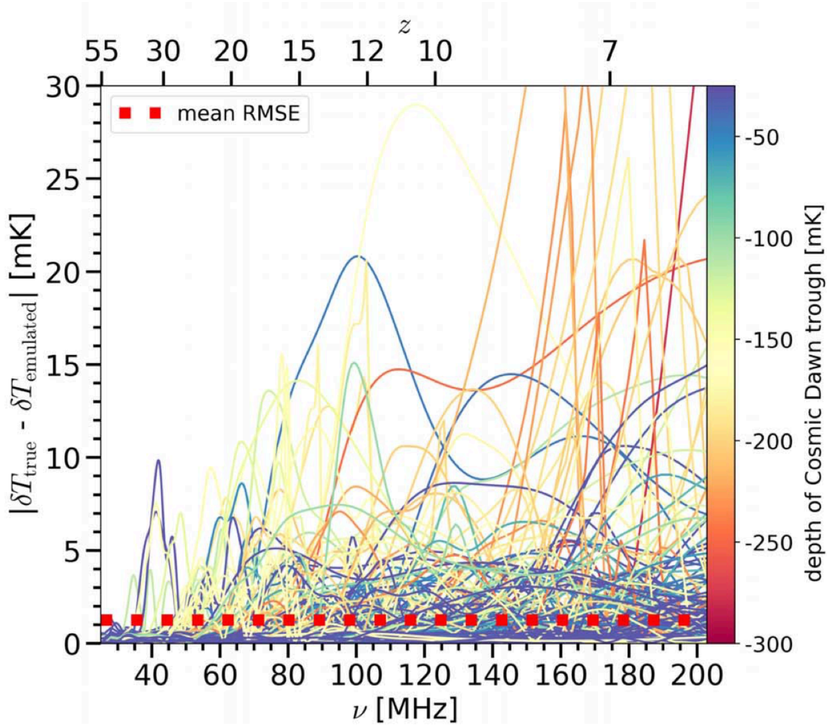
Metric	With AFB, with resampling	Resampling only	AFB only	No AFB, no resampling	Dorigo Jones et al. (2023) No AFB, no resampling
Mean	**0.82**	1.27	1.34	1.05	1.25
95^{th} Percentile	**2.56**	2.97	4.48	3.31	—
Worst	18.91	**14.54**	33.28	26.42	18.5

globalemu performance and ARES modelling



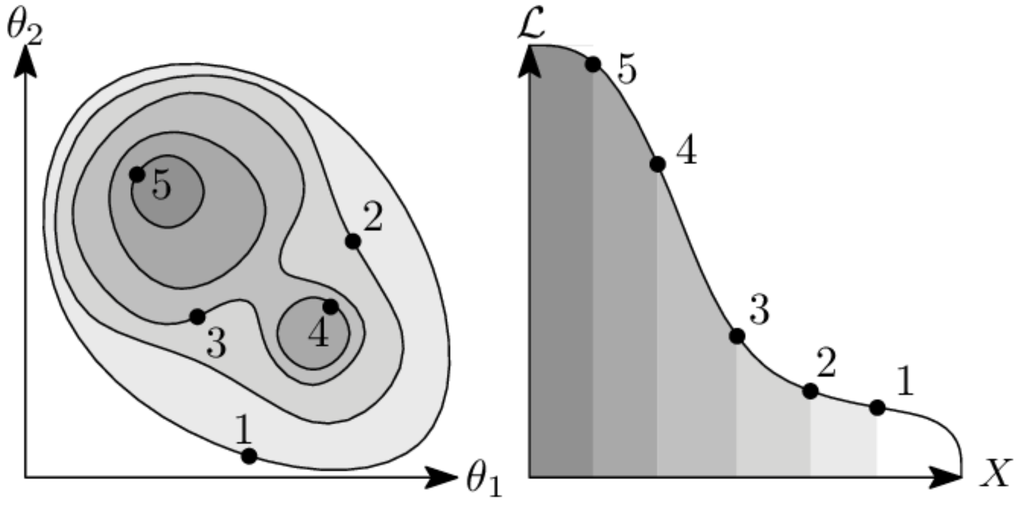
A side note on Dorigo Jones+23 emulator

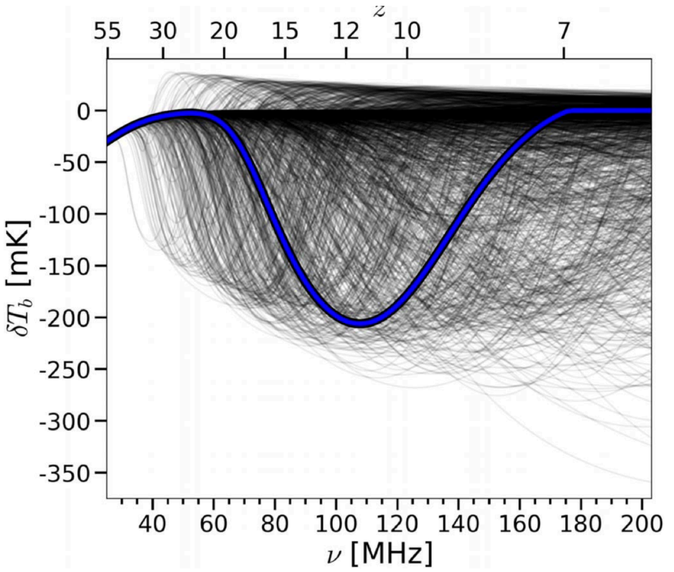




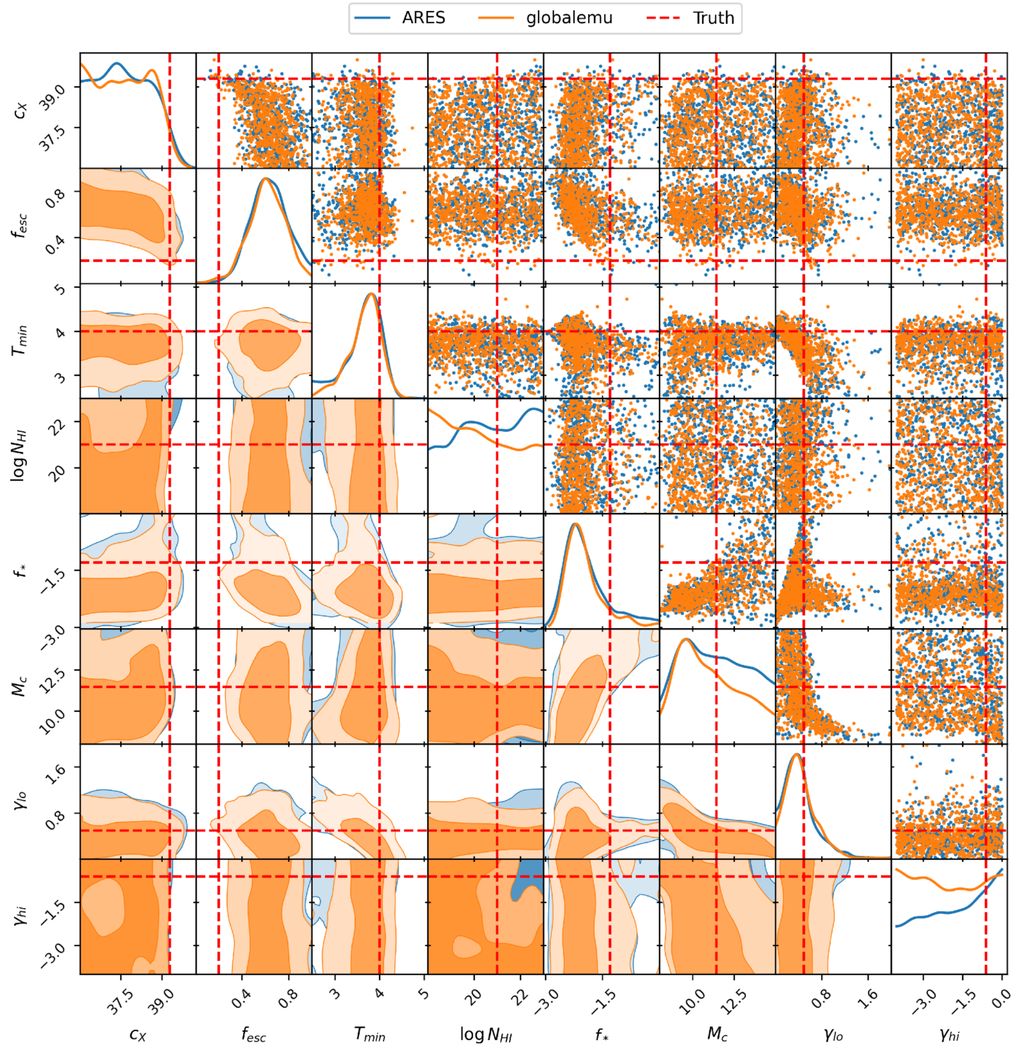
Running the analysis

- Same fiducial signal as in Dorigo Jones+23
- Same prior range and same sampler
- No UVLF
- Assuming as Gaussian likelihood as was done in their paper
- Assuming absolute knowledge of the level of noise in the data (not really a good idea)
- Running for 5, 25 and 50 mK

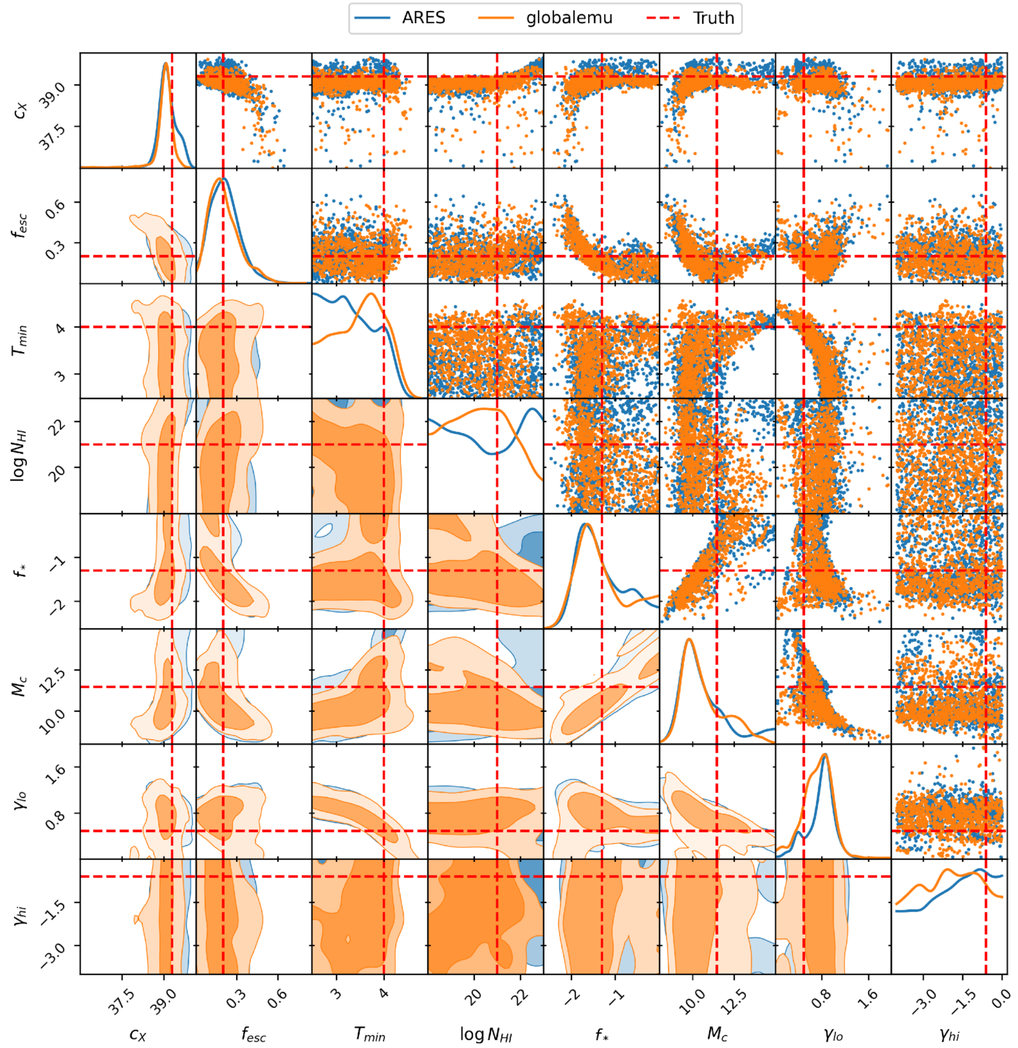




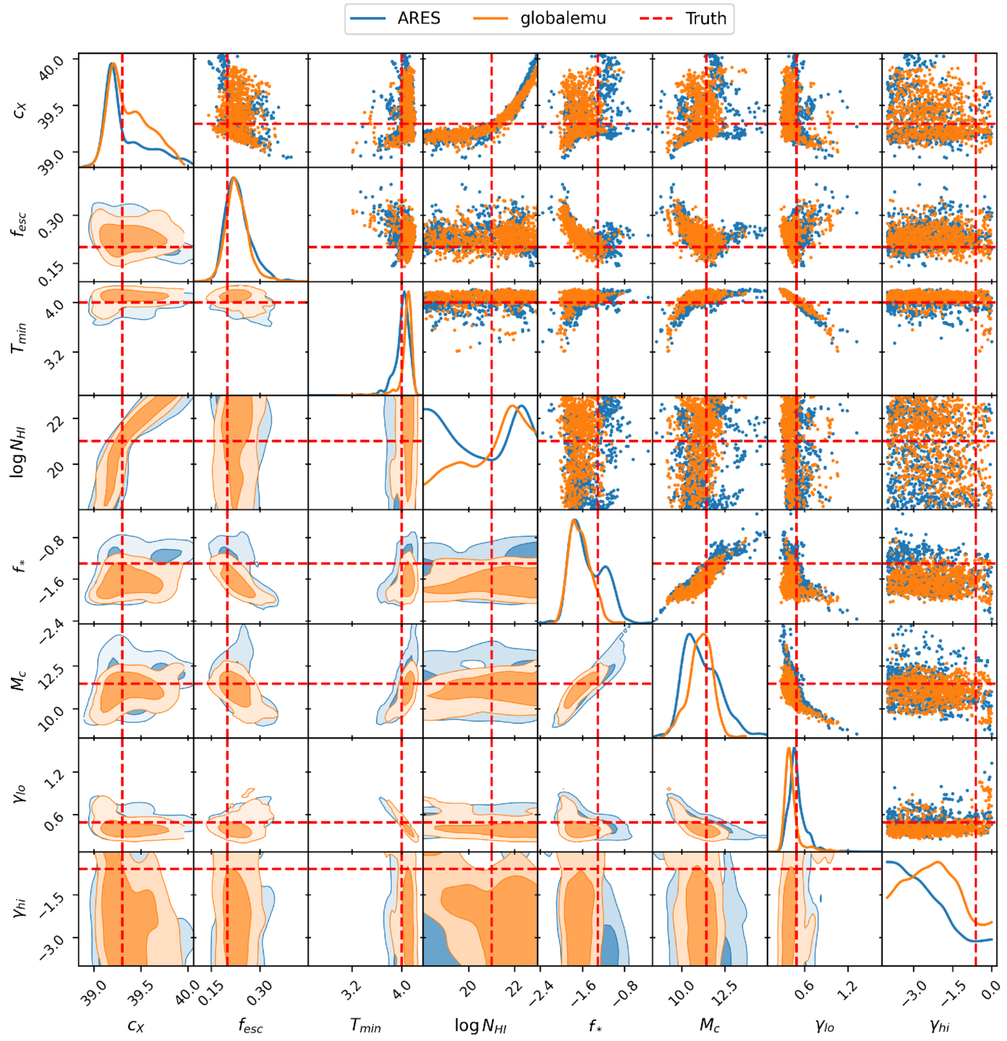
Running the analysis - 50 mK



Running the analysis - 25 mK

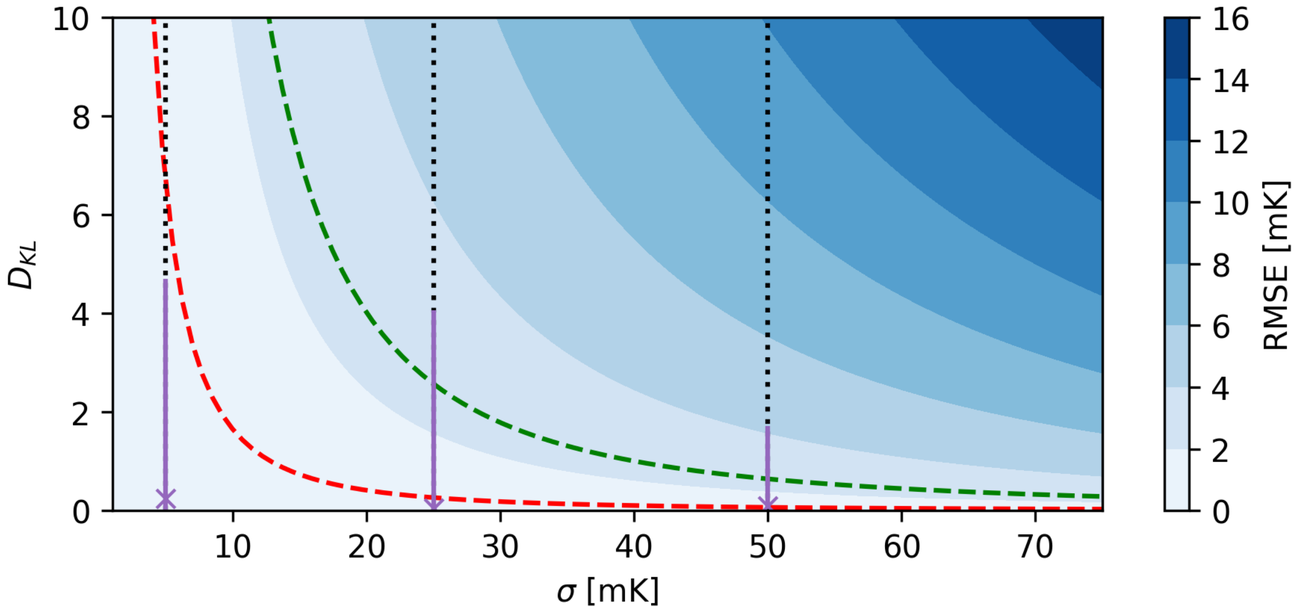


Running the analysis - 5 mK



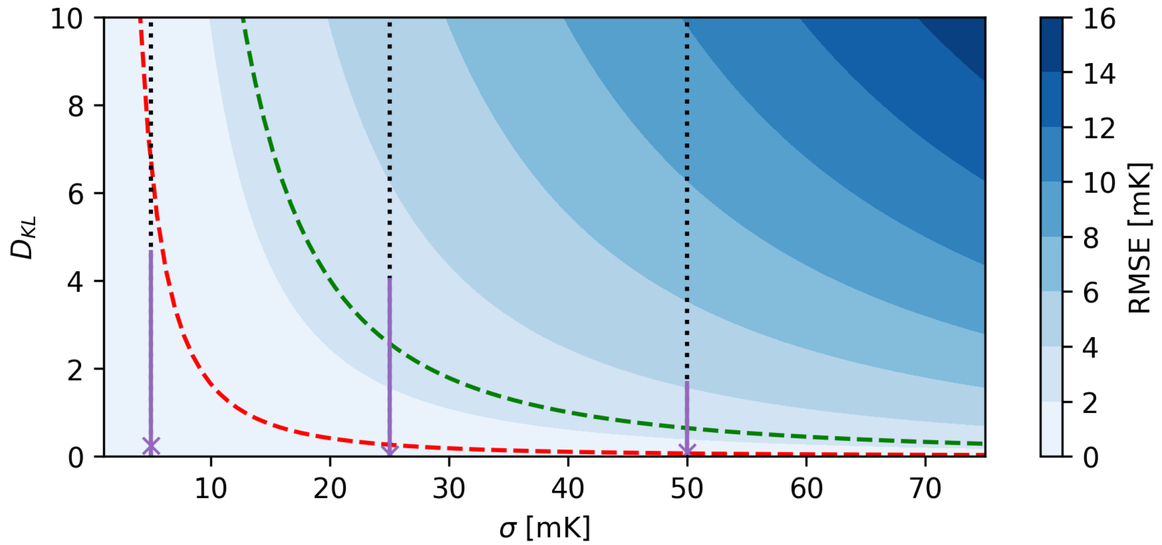
How about the D_{KL}?

- Need to be able to evaluate the log-probability for sets of samples on both distributions to get D_{KL}
- Use normalising flows implemented with margarine
- Compare D_{KL} with predicted upper limits

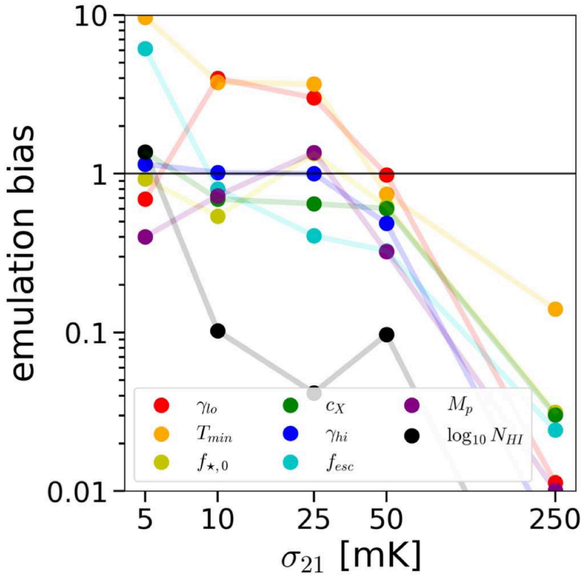


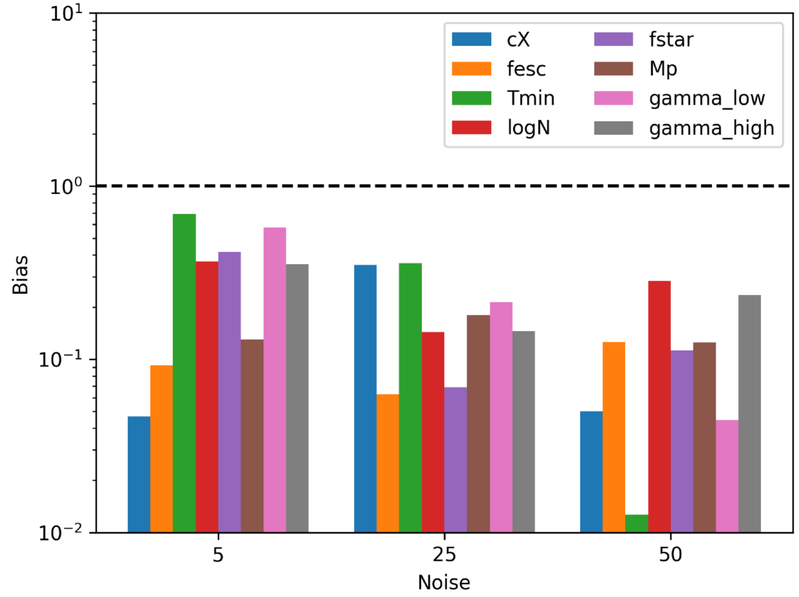
Comparison with Dorigo Jones + 23

Comparing our conclusions



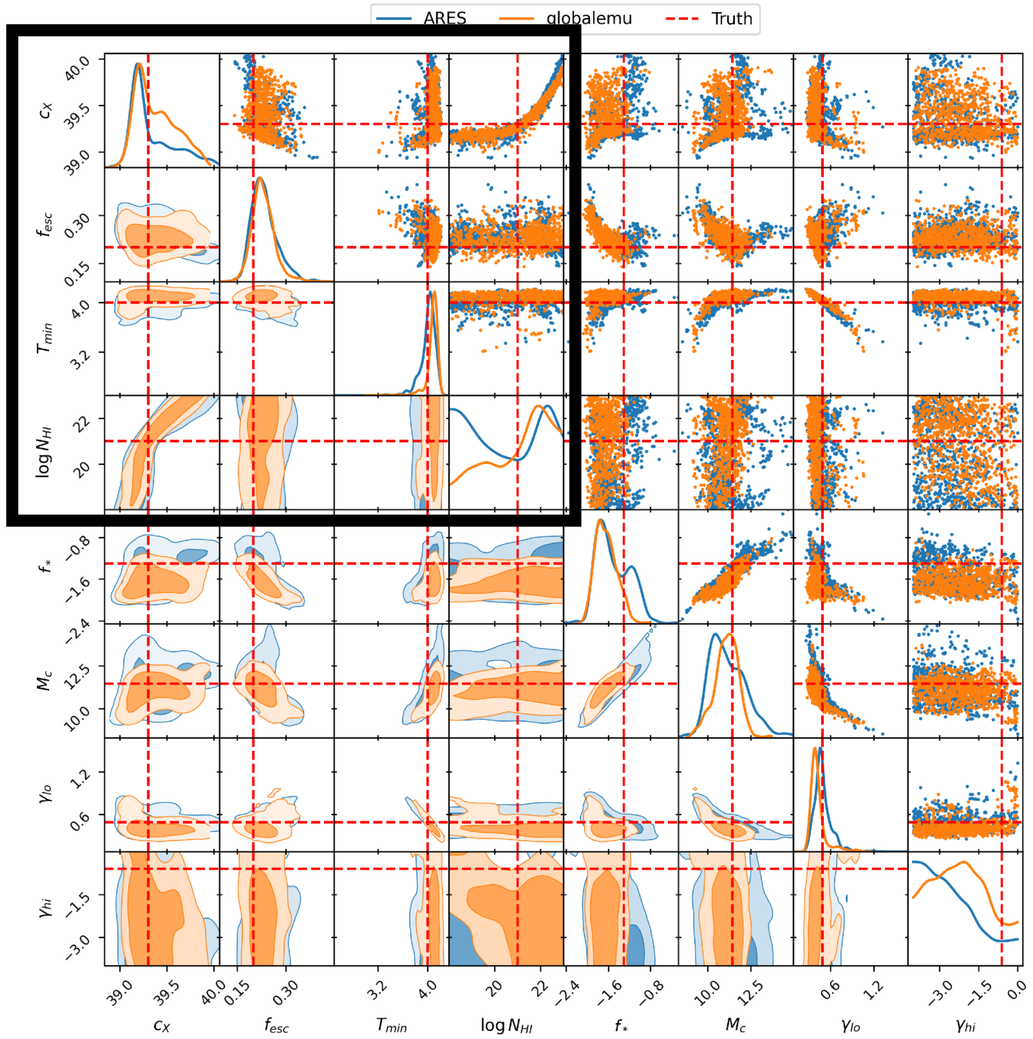
Noise Level	Predicted $D_{\mathrm{KL}} \leq$		Actual D_{KL}
	Mean RMSE	95th Percentile	
5	6.59	64.23	$0.25^{+4.45}_{-0.25}$
25	0.26	2.57	$0.05^{+4.02}_{-0.52}$
50	0.06	0.64	$0.09^{+1.62}_{-0.03}$



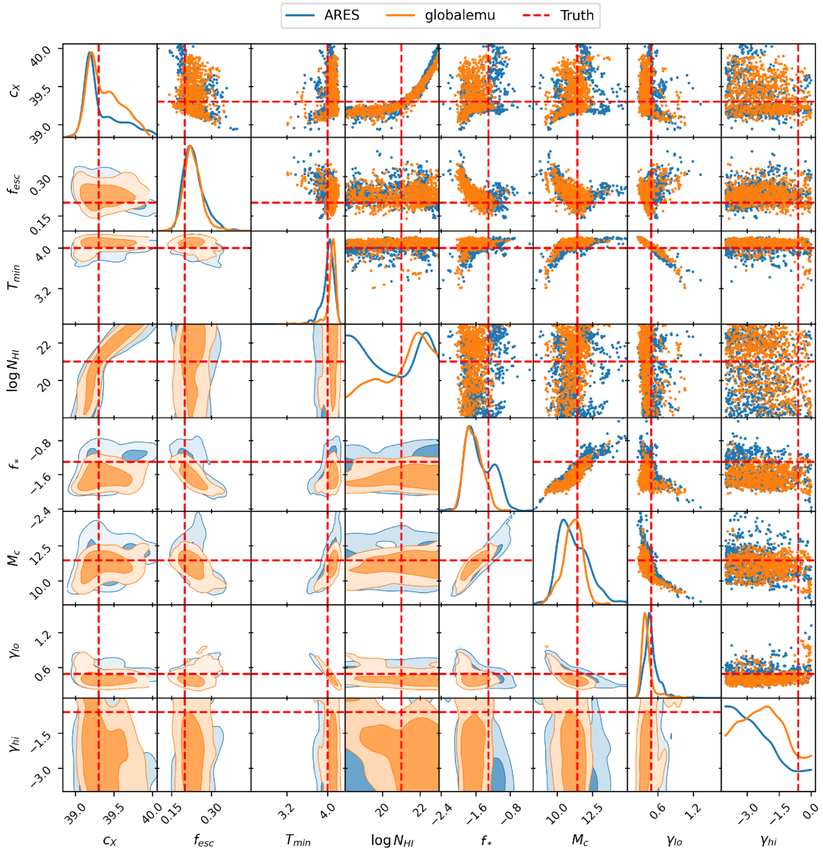


Comparing our Posteriors

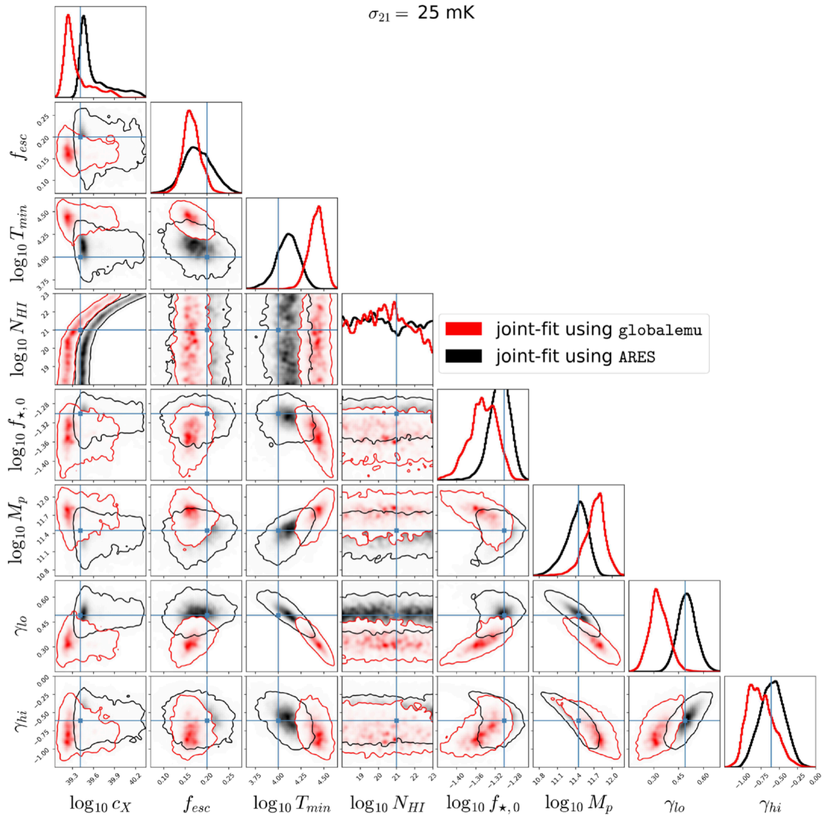
- Don’t pay too much attention to the differences in $f_*, M_c(\text{their } \mathrm{M_p}), \gamma_{lo}, \gamma_{hi}$ because these parameters are constrained by UV luminosity functions
- We expect our true posteriors to look very similar to theirs for the parameters constrained by the 21cm signal e.g. $f_{\rm esc}, \log N_{HI}, C_X, T_{\rm min}$ but this isn’t what I see
- Effectively ignore the emulated posteriors here
- Just focusing on a comparison of the true posteriors
- That’s model ARES with ARES

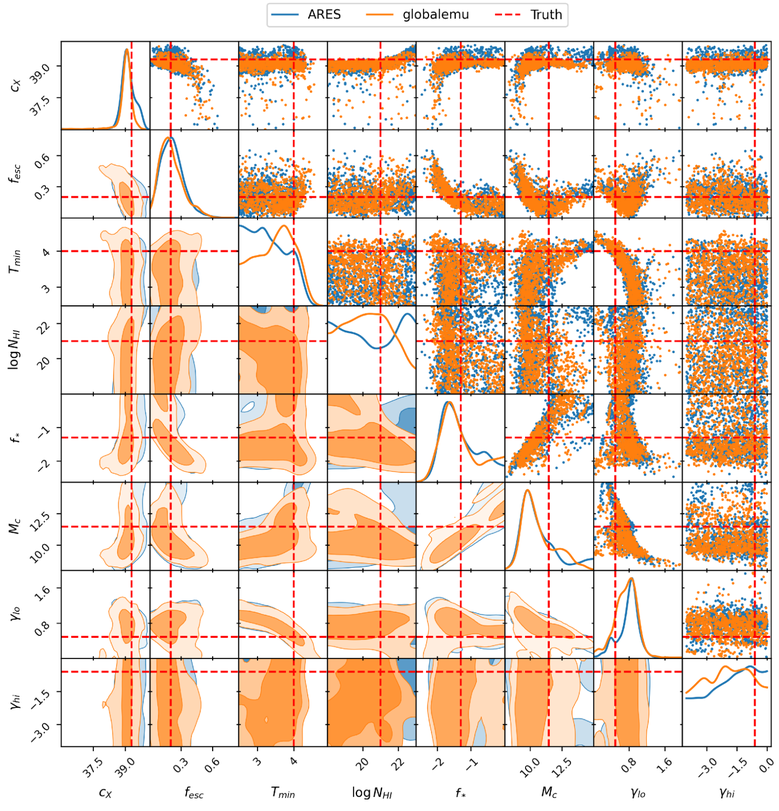


Comparing blue and black here…

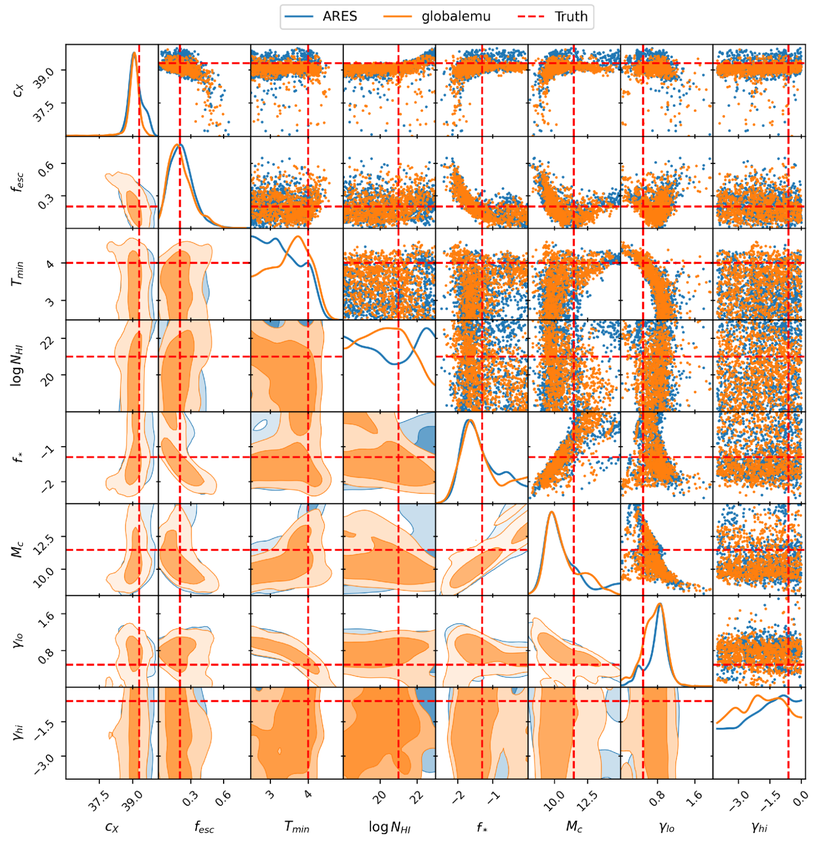


Our 5 mk

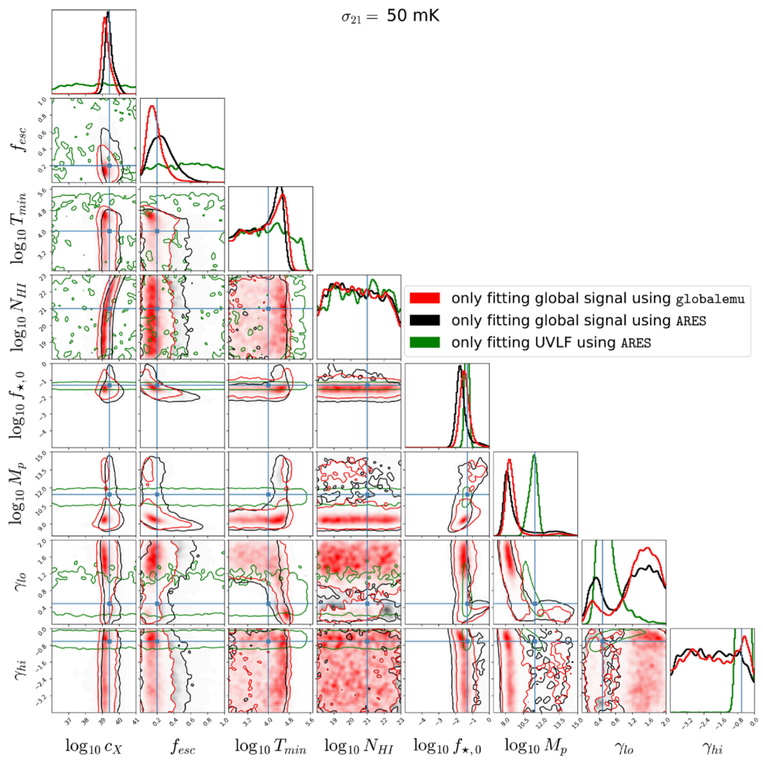


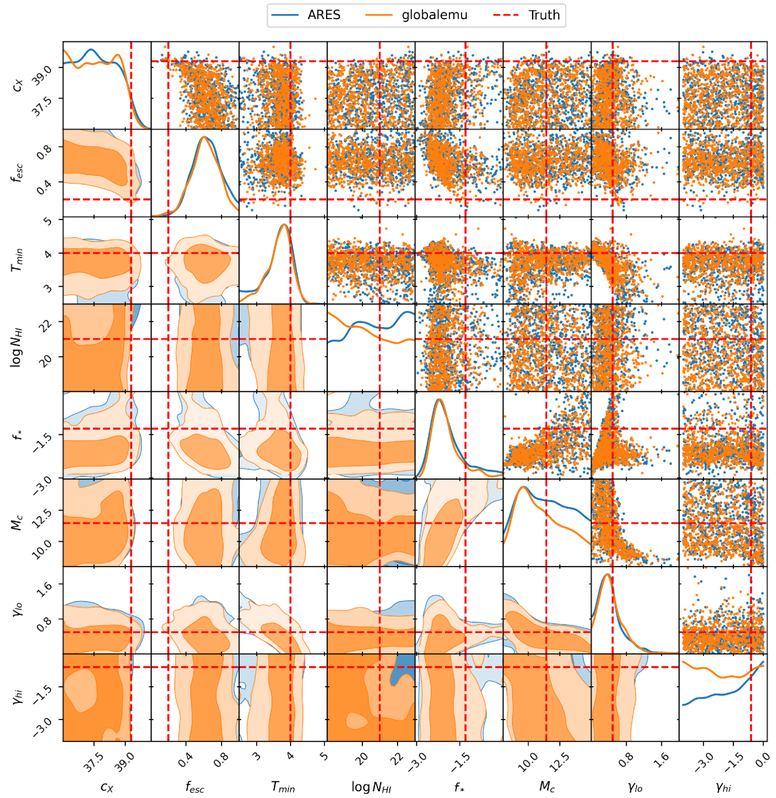


Our 25mK

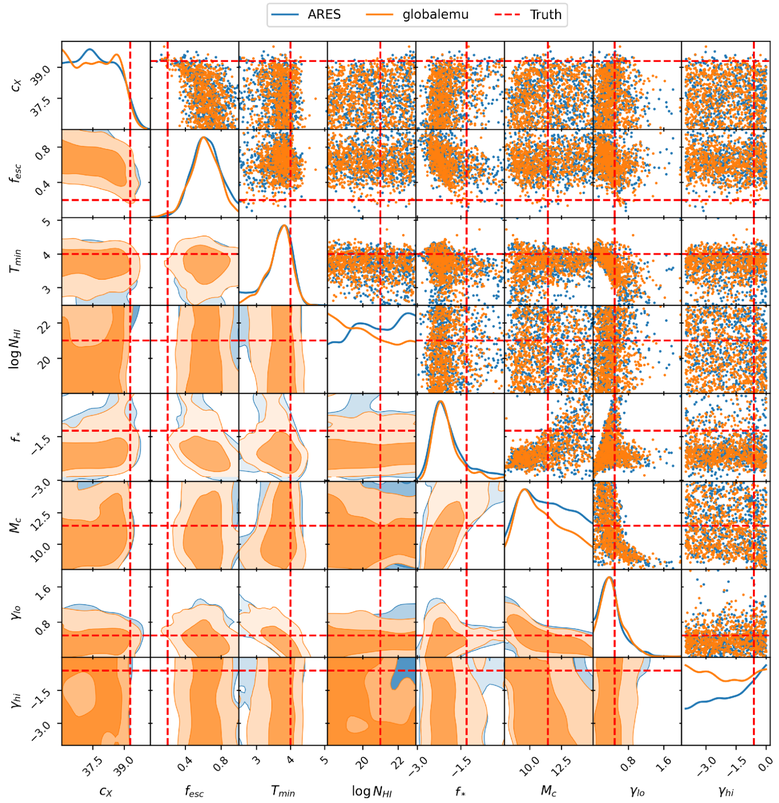


Our 25mK

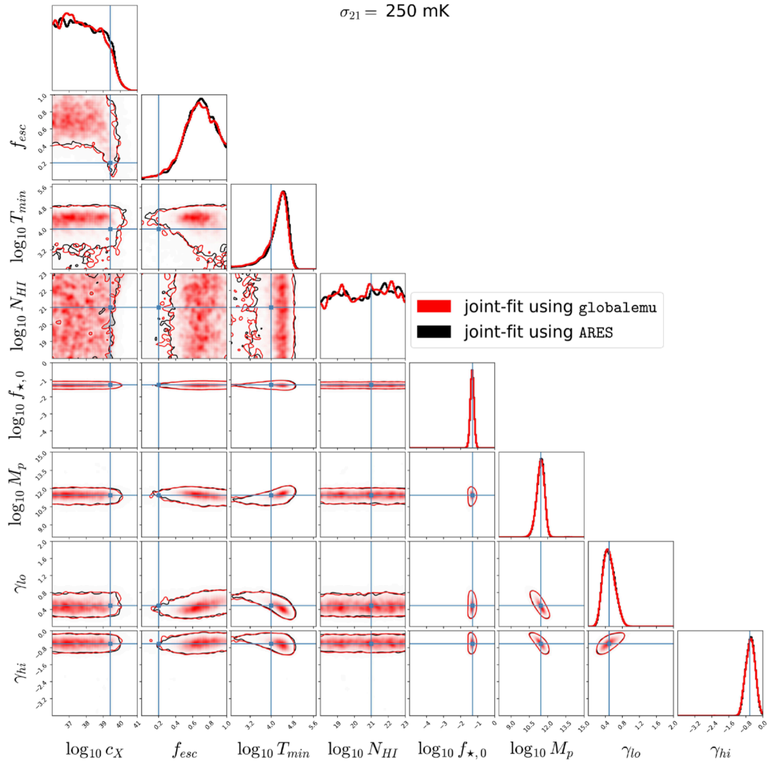




Our 50 mK

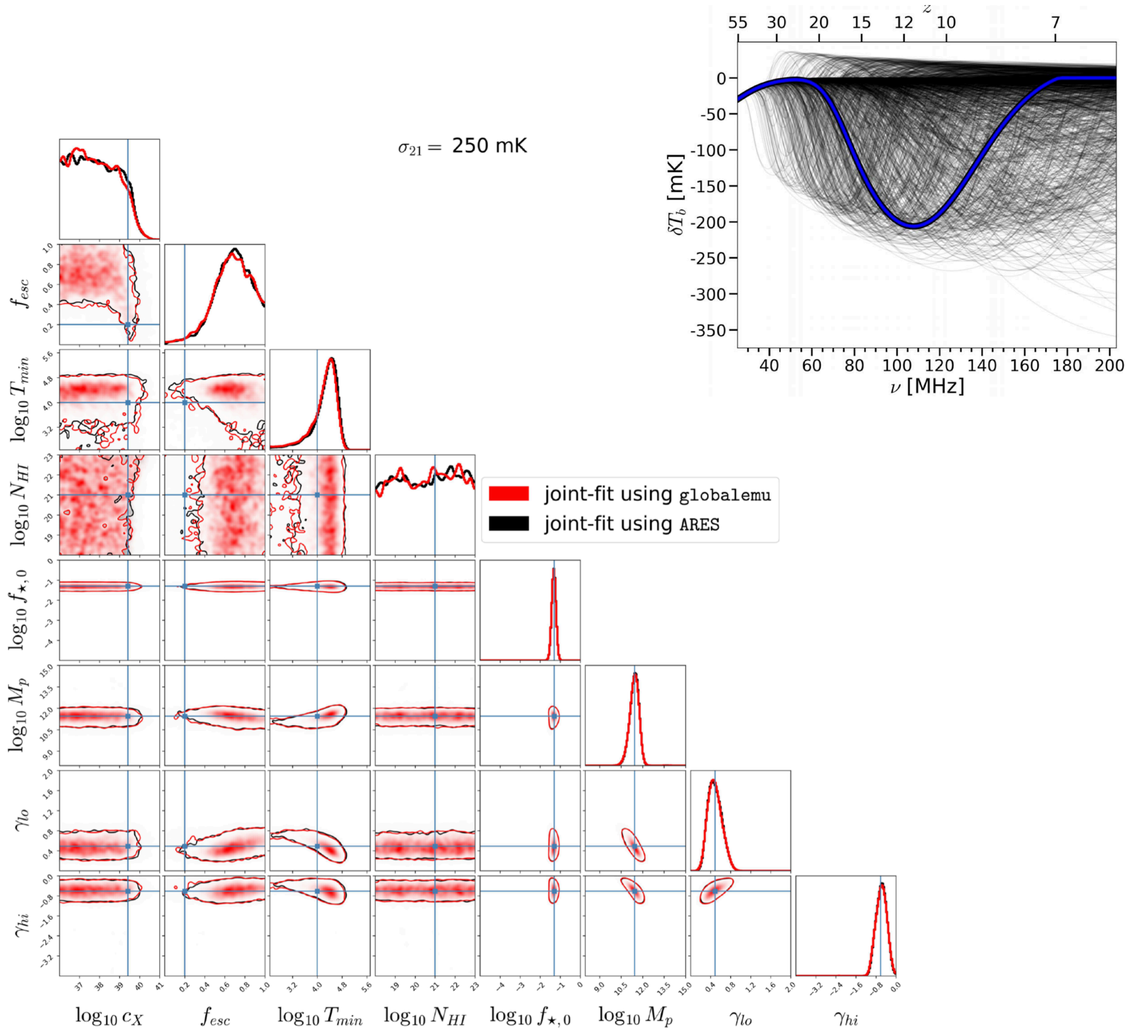

ARES
globalemu
Truth
c_X
f_{esc}
T_{min}
$\log N_{HI}$
$f_\star$
M_c
γ_{lo}
γ_{hi}

Our 50 mK


$\sigma_{21} =$ 250 mK
joint-fit using globalemu
joint-fit using ARES
$\log_{10} c_X$
f_{esc}
$\log_{10} T_{min}$
$\log_{10} N_{HI}$
$\log_{10} f_{\star,0}$
$\log_{10} M_p$
γ_{lo}
γ_{hi}

SNR < 1?

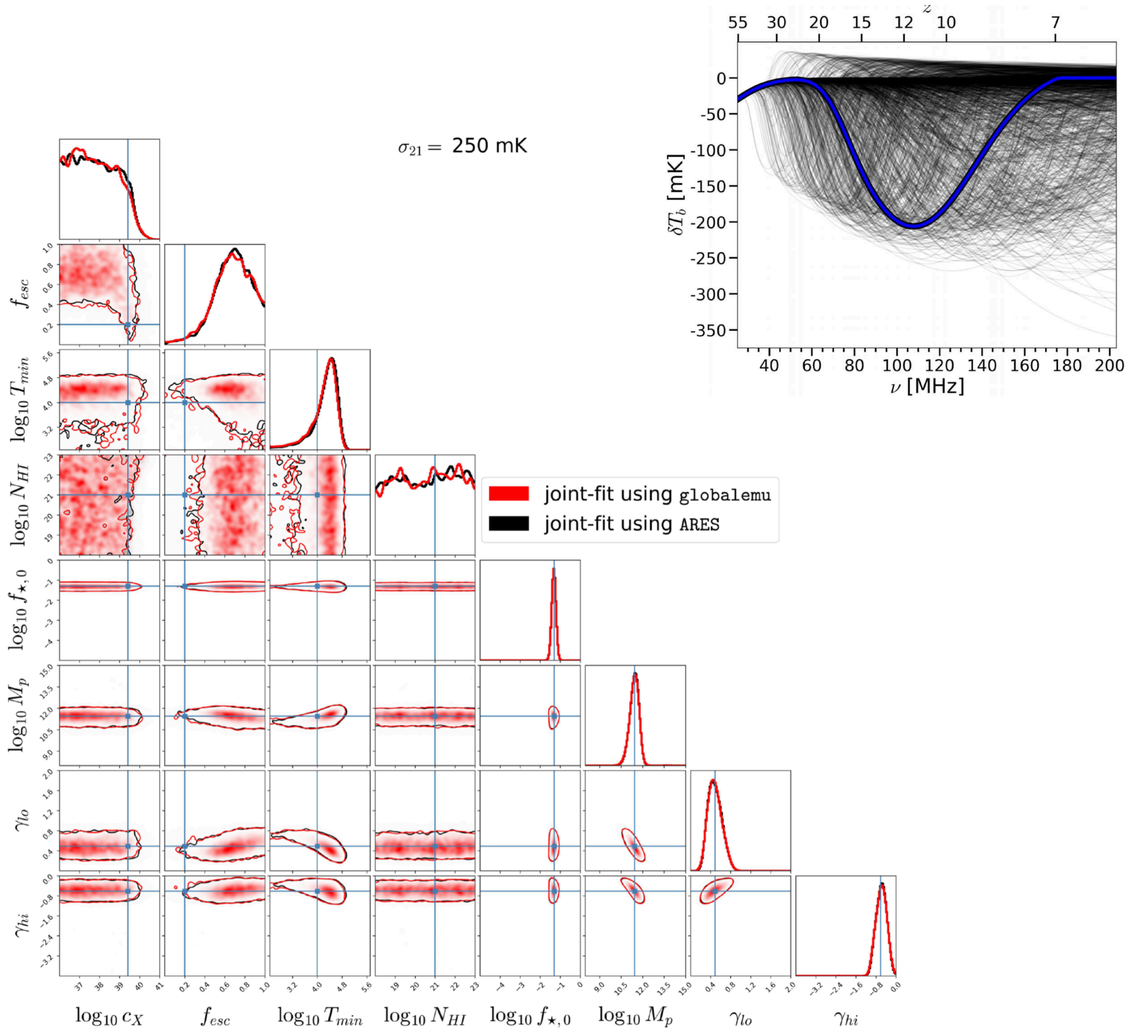
- The fiducial 21cm signal has a maximum depth of around 200 mK
- If this truly is for a $\sigma = 250$ mK then the SNR is at best 1
- They really should not see constraints in parameters like $f_{\rm esc}$ and $T_{\rm min}$



Conclusions

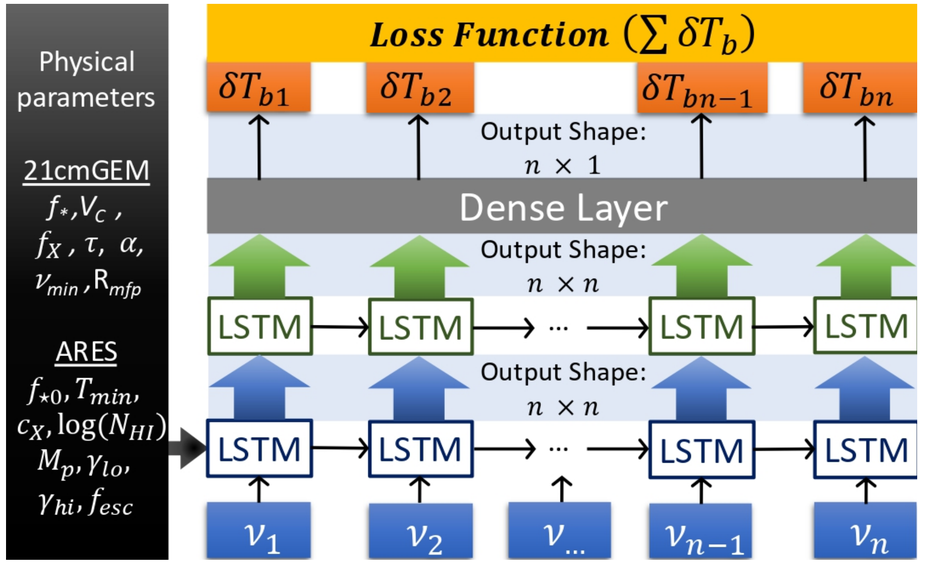
Conclusions - Dorigo Jones+23?

- Something went wrong in Dorigo Jones+23
- Hard to tell what because they did not make their code public and they complicated their work with UVLF constraints
- Looks like they even mislabelled graphs
- Even though there are some minor differences in the emulators we'd still expect much better posterior recovery than they saw



Conclusions - Dorigo Jones+24?

- Dorigo Jones and collaborators recently introduced a new emulator for the 21cm signal
- Using Long-Short term memory networks
- They claimed “emulation error < 1 mK is needed to sufficiently exploit optimistic or standard measurements of the 21cm signal and obtain unbiased posteriors”
- Did not repeat the analysis in Dorigo Jones+23 to justify that their new emulator was good enough



Conclusions - A more positive note?

- We are presenting a useful upper bound on the incurred information loss from using emulators in inference
- Broadly applicable beyond 21cm
- We demonstrated that we can accurately recover posteriors even with $\bar{\epsilon} \approx 0.2\sigma$
- We are reestablishing confidence in what is a really important tool in the field
- Paper and code soon