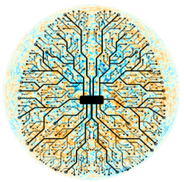
On the accuracy of posterior recovery with neural networks

Harry Bevins
Kavli Fellow @ Kavli Institute for Cosmology, Cambridge
with
Thomas Gessey-Jones, Will Handley


The Importance of emulators

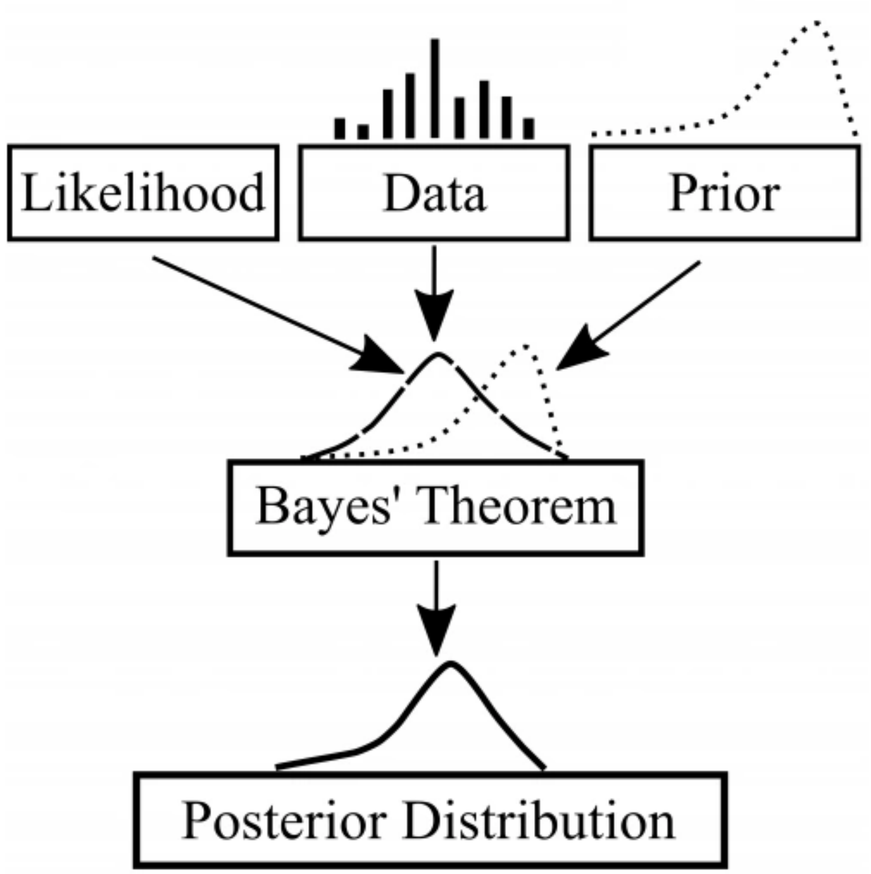
Inference in Cosmology

- To fully understand modelling uncertainties and degeneracy between model parameters we use Bayesian inference

$$P(\theta|D,M) = \frac{P(D|\theta,M)}{P(D|M)}P(\theta|M)$$

$$PZ = L\pi$$

where D is our data, θ are our model parameters and M is our model



Strong dependence on model runtime

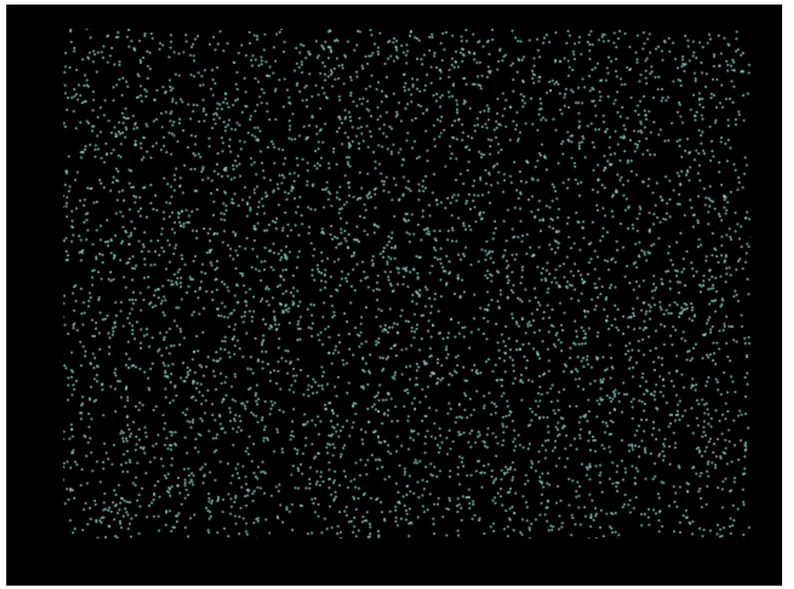
Nested (general) sampling runtime

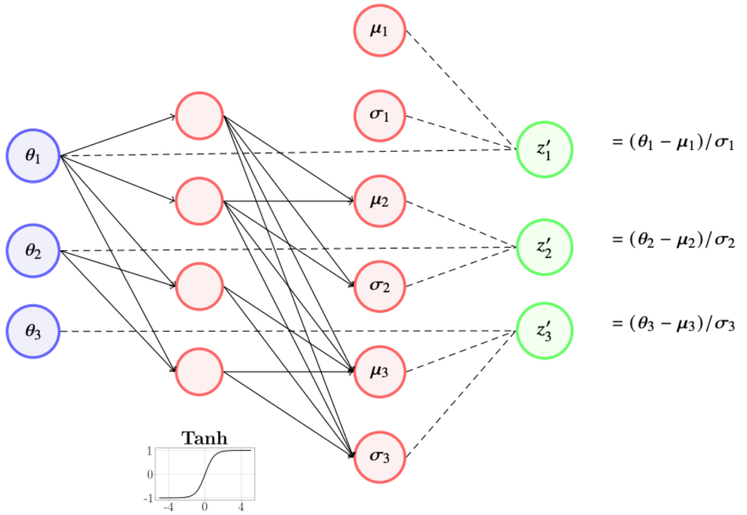
$$T \propto n_{\text{live}} \times \langle T\{L(\theta)\}\rangle \times \langle T\{\text{impl}\}\rangle \times D_{\text{KL}}(P\,||\,\pi)$$

$\langle T\{\text{impl}\}\rangle \rightarrow$ **BlackJAX NS (Yallup et al. 2025)**

$D_{KL}(P\,||\,\pi) \rightarrow$ **Better priors with normalising flows using margarine (Bevins et al 2022, 2023)**

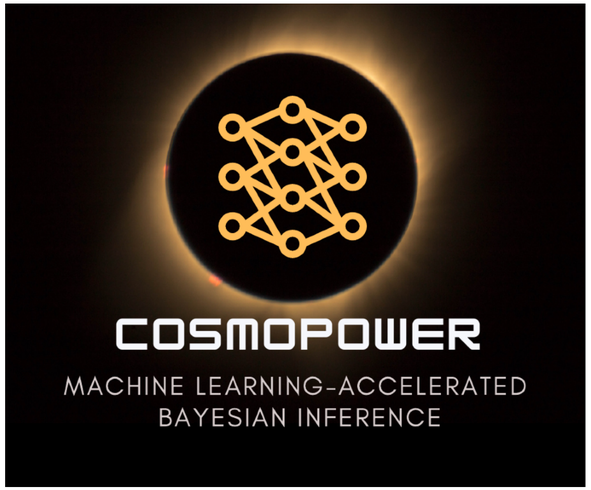
$\langle T\{L(\theta)\}\rangle \rightarrow$ **Emulators**

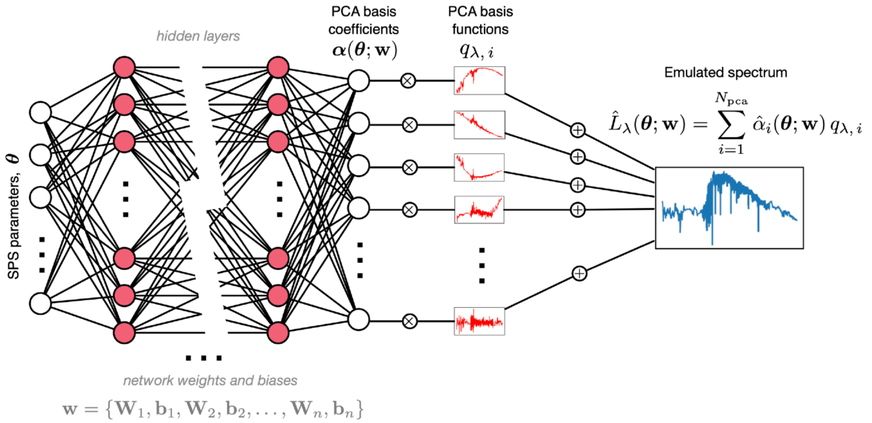




Emulators in Cosmology

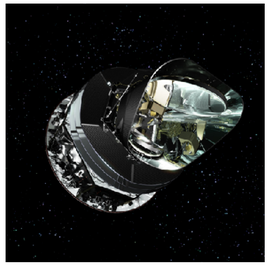
- Neural network emulators are really important in Cosmology and Astrophysics
- Modelling gravitational waves, CMB observables, galaxy SEDs, 21-cm signal etc is very complicated!
- For fast inference on computationally expensive likelihoods
- Generating large training data sets for training simulation based inference algorithms

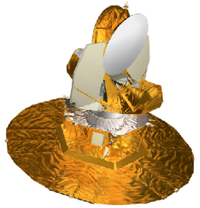


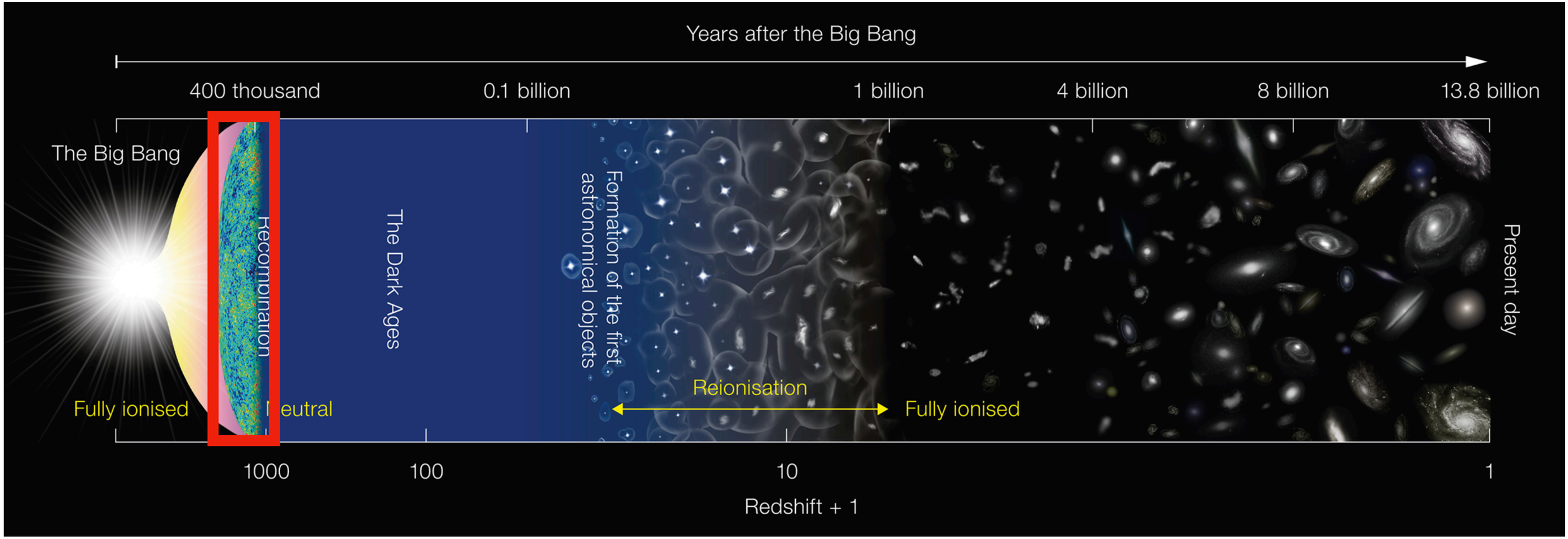


The Infant Universe and 21-cm Cosmology

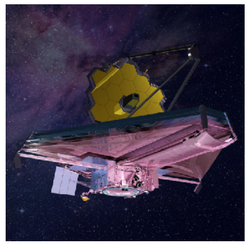
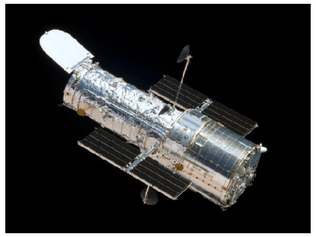
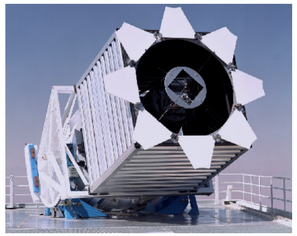
A brief history of the universe

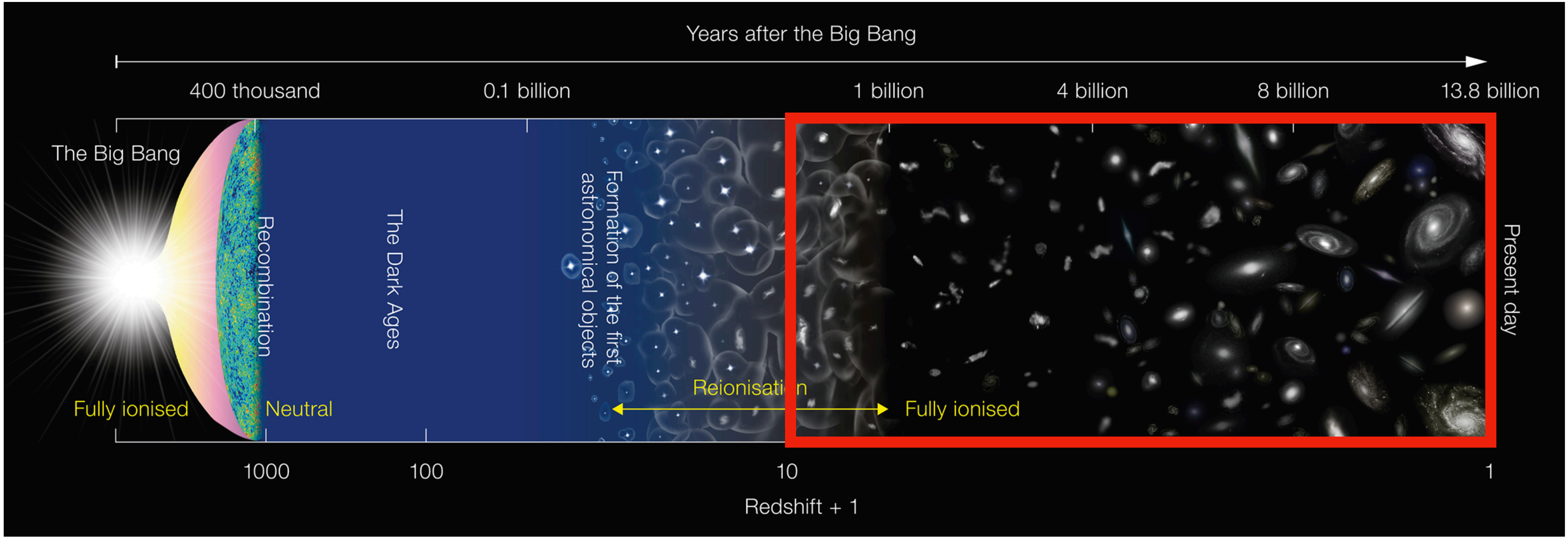




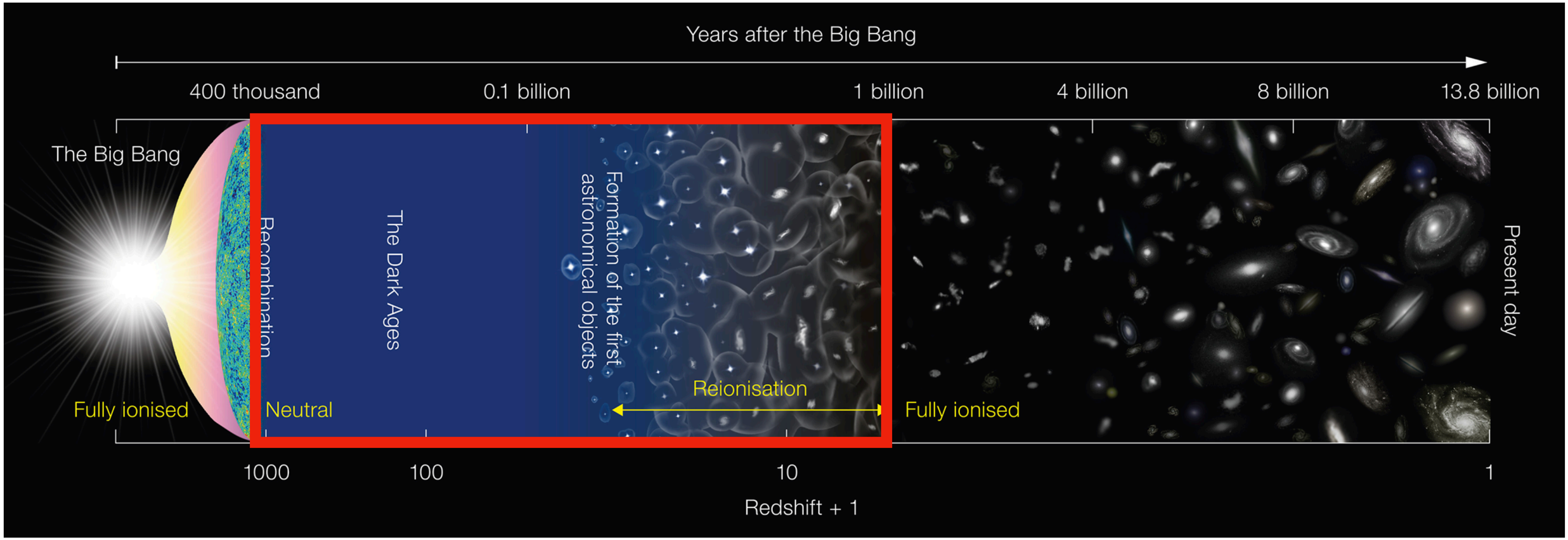


A brief history of the universe

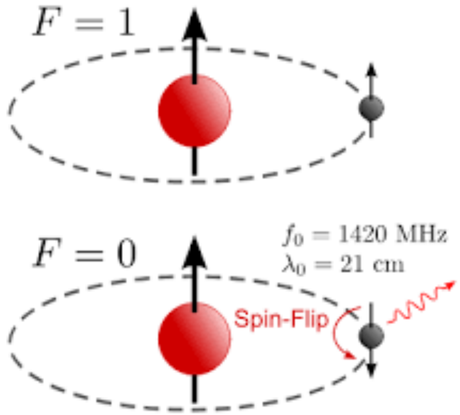


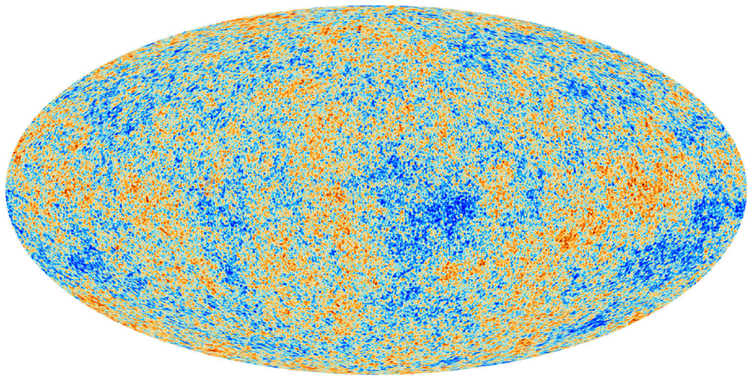
A brief history of the universe



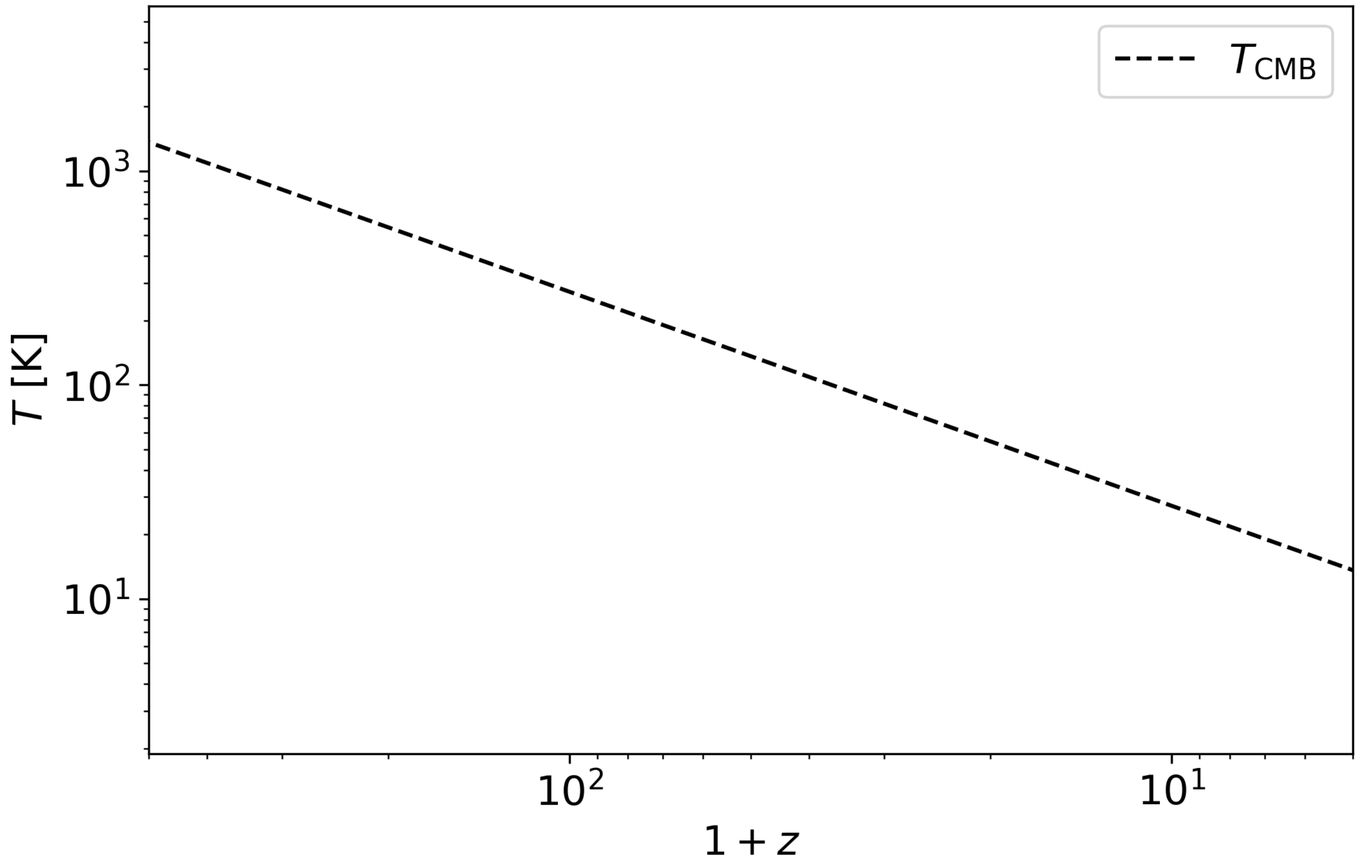
21-cm Cosmology

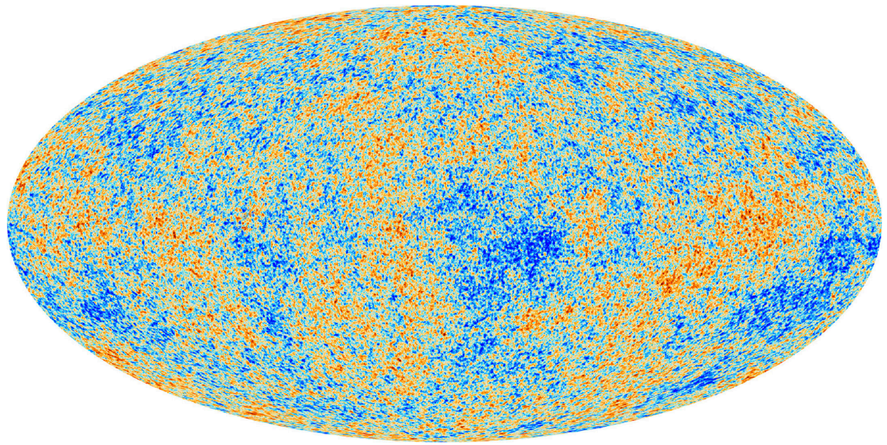
- Spin-flip transition in neutral hydrogen
- Forbidden transition that can't be seen in the lab
- Define the spin temperature
- Measure relative to the radio background
- To understand the importance of the spin temperature we look at the thermal history of the universe



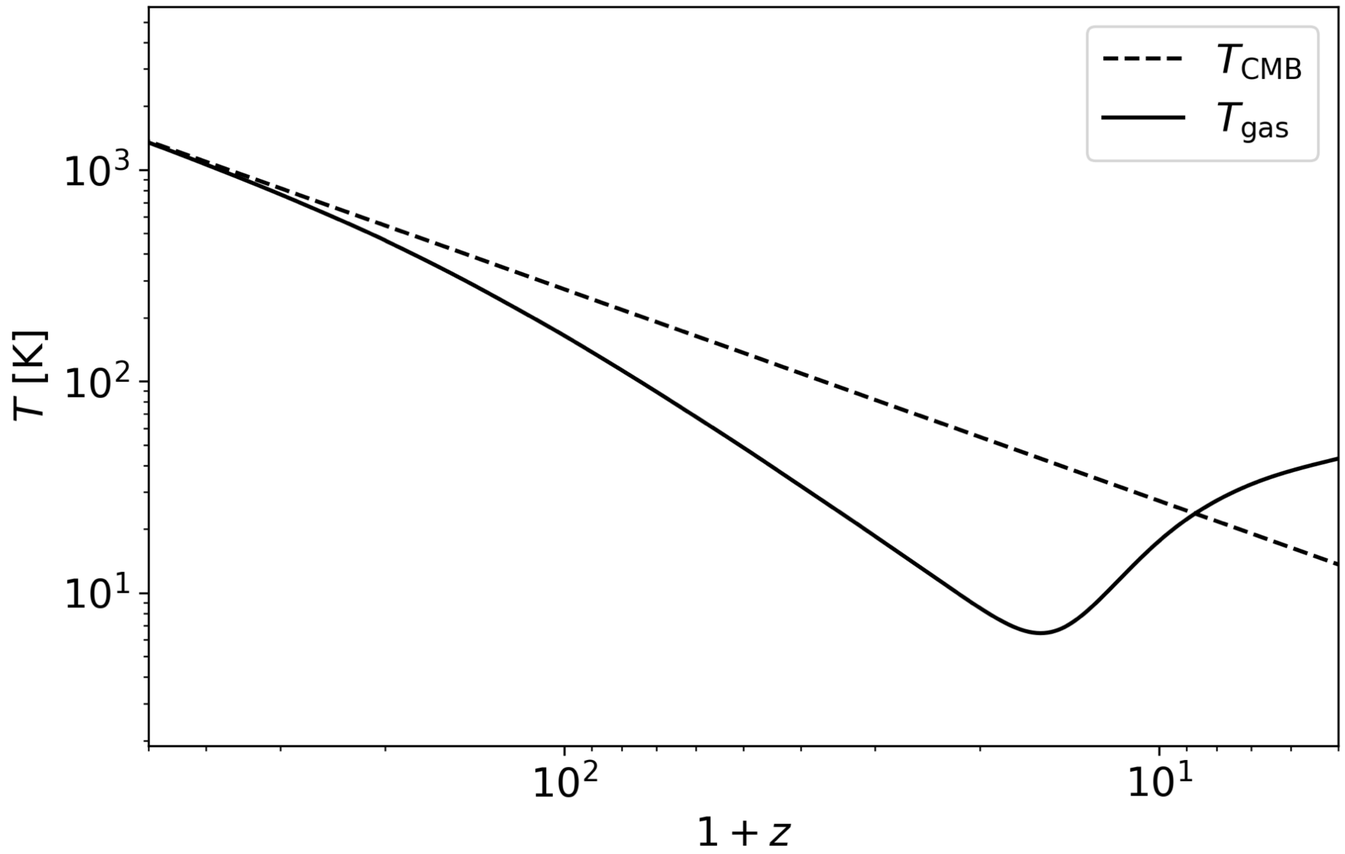


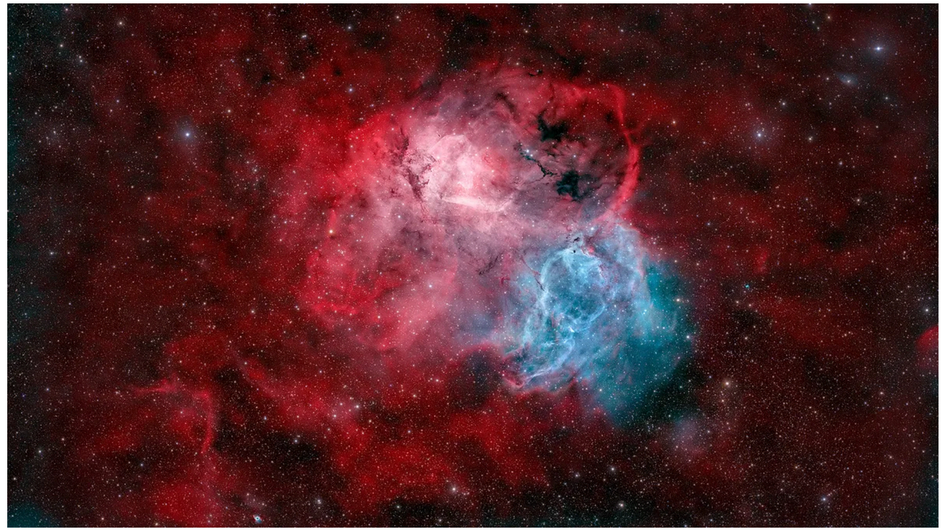
21-cm Cosmology



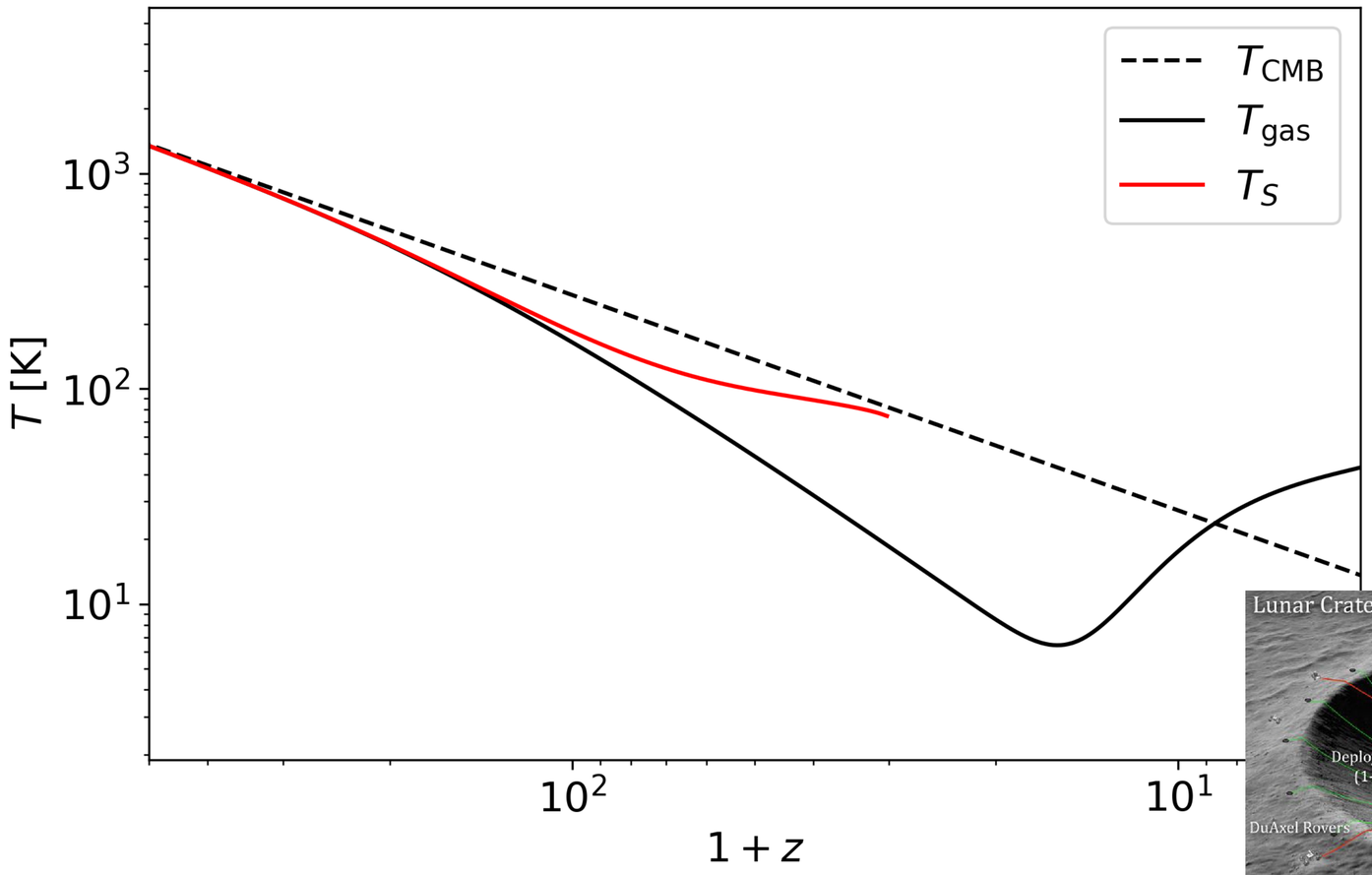


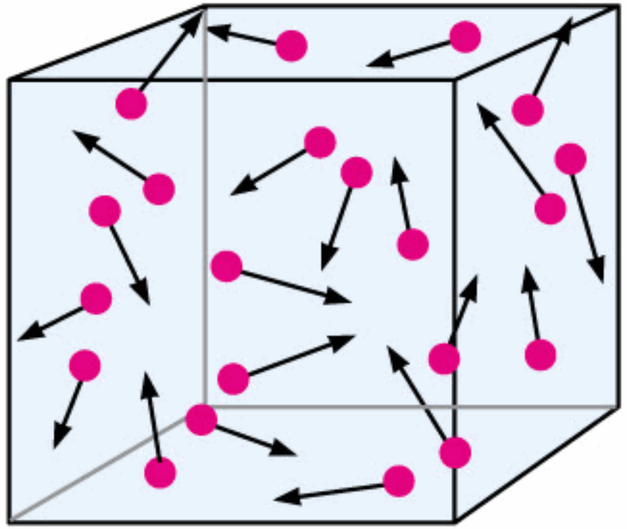
21-cm Cosmology

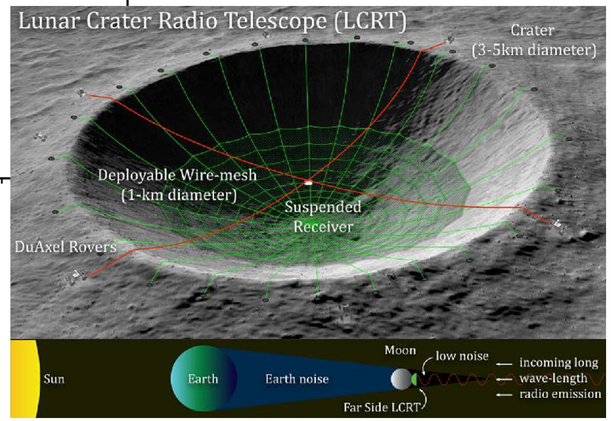


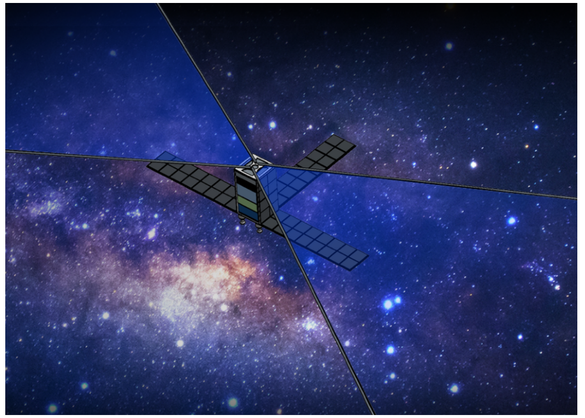


Dark Ages ($z \lesssim 30; t \lesssim 0.5$ Gyr)



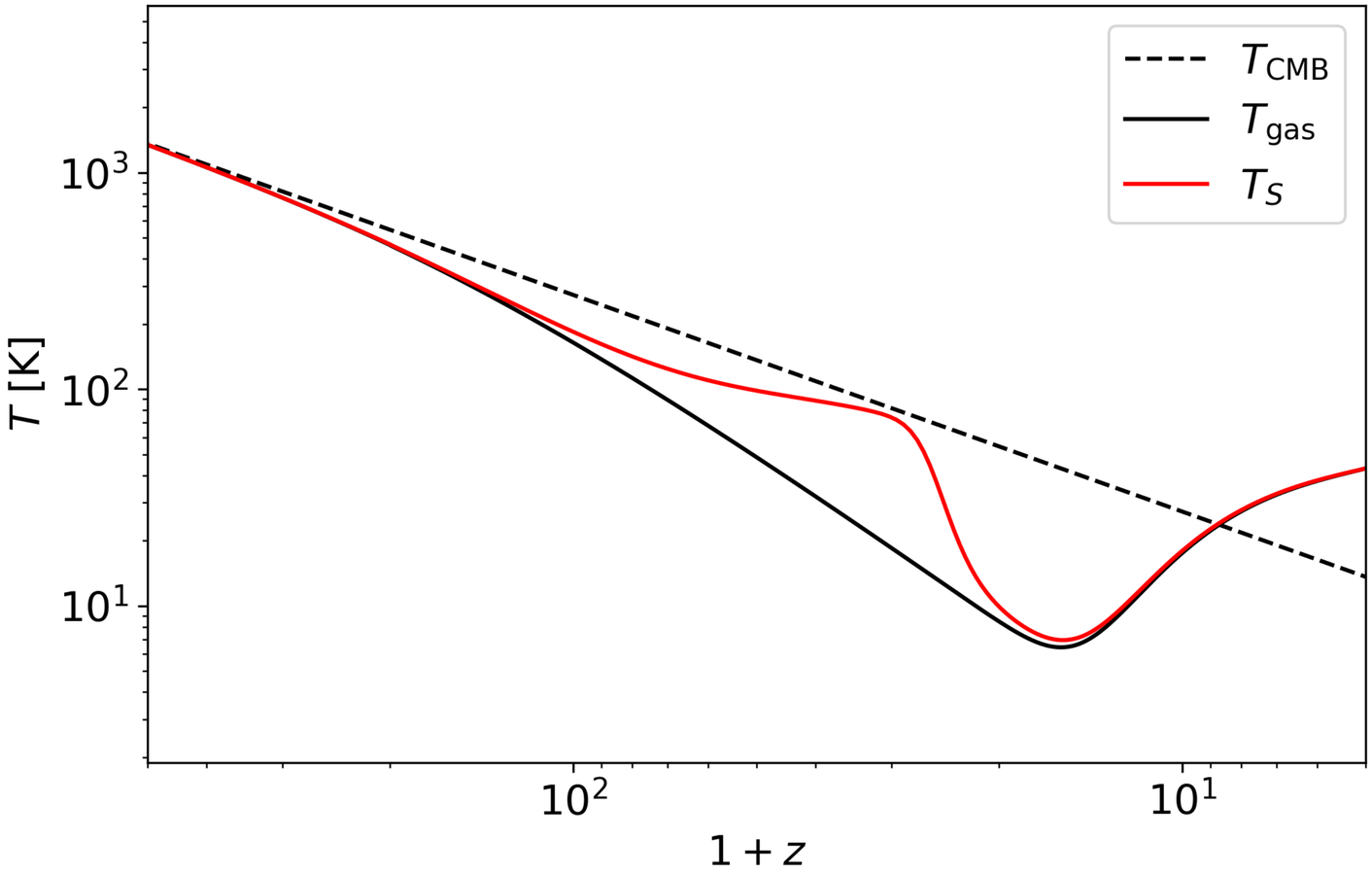


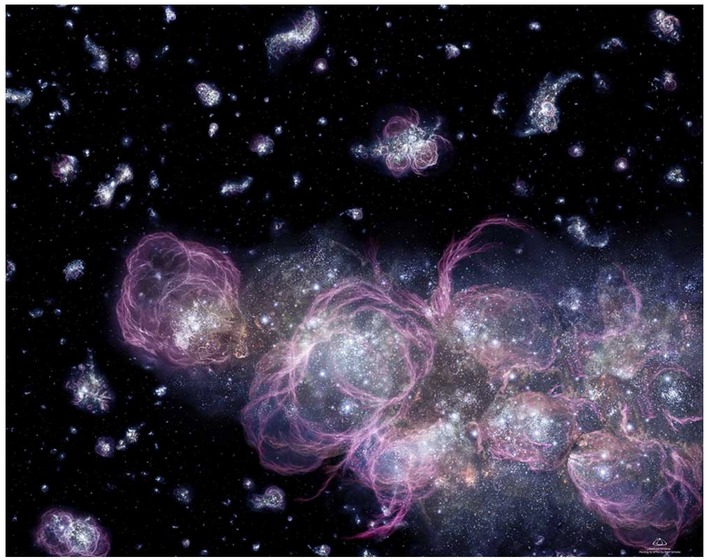


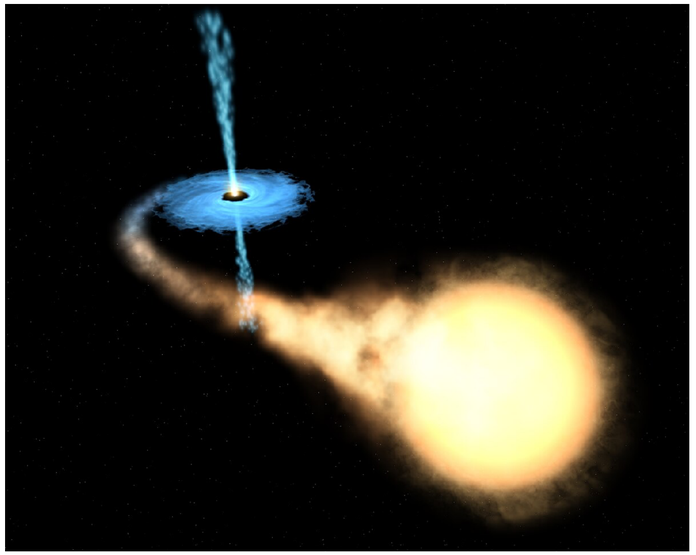


Cosmic Dawn and Reionisation

$(30 \lesssim z \lesssim 5; t = 0.5 - 12.5\ \mathrm{Gyr})$

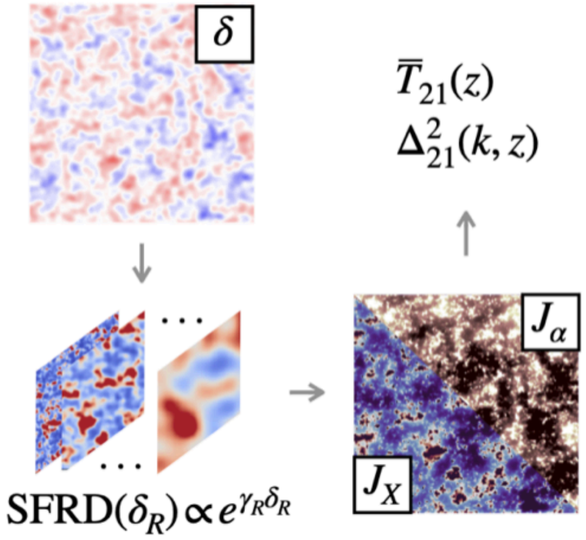


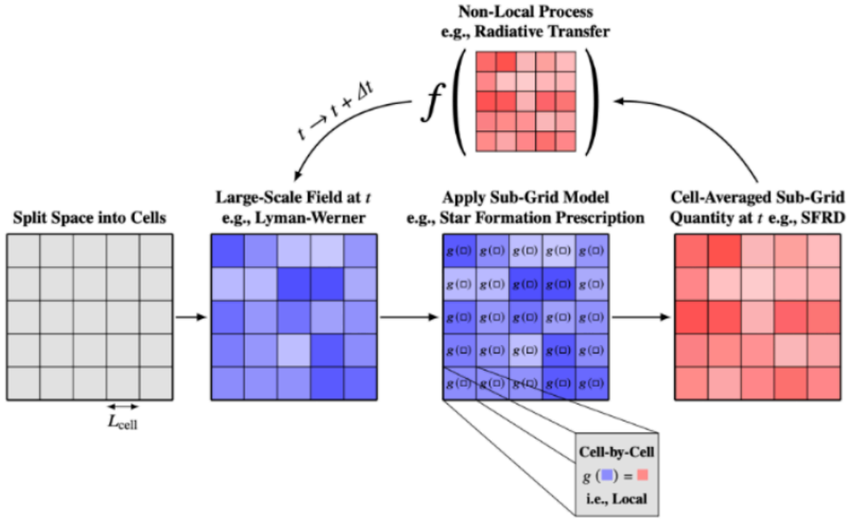




Simulating the 21-cm signal

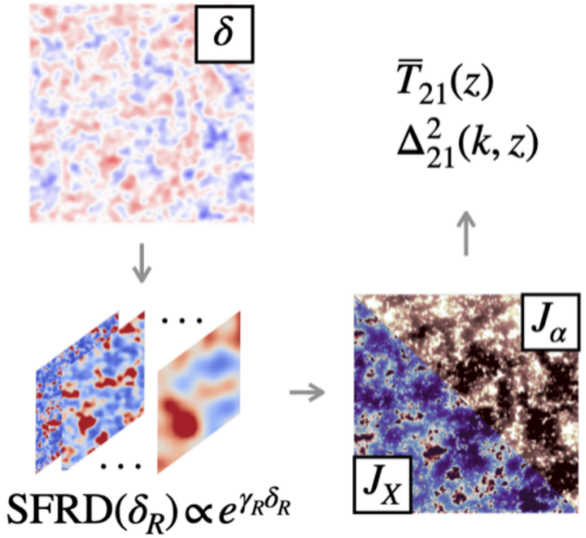
- Several classes of simulation
  - Analytic models (Zeus21, Echo21…)
    - Make approximation
    - Evaluate in seconds
  - Semi-numerical models (21cmSPACE, 21cmFAST…)
    - Detailed grid model populated with halos and galaxies
    - Evaluate in hours
  - Hydrodynamical codes (C2-RAY…)
    - Radiative transfer codes with hydrodynamics and feedback
    - Evaluate in days to weeks

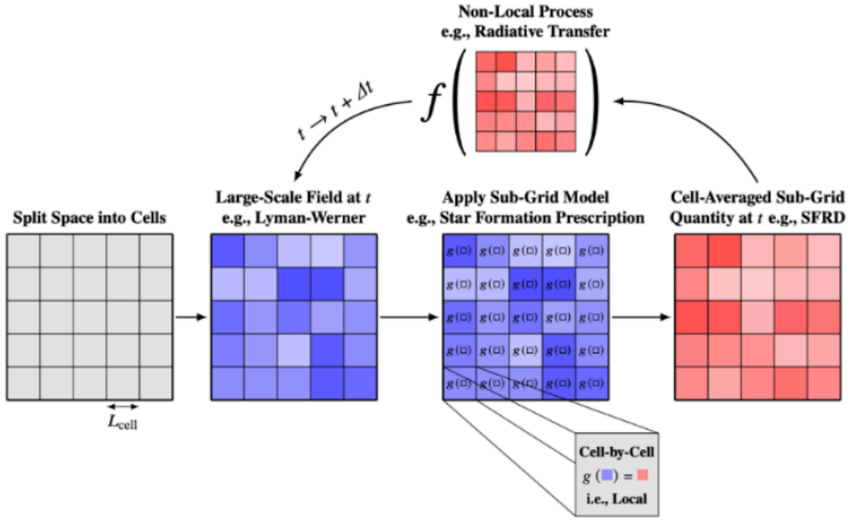




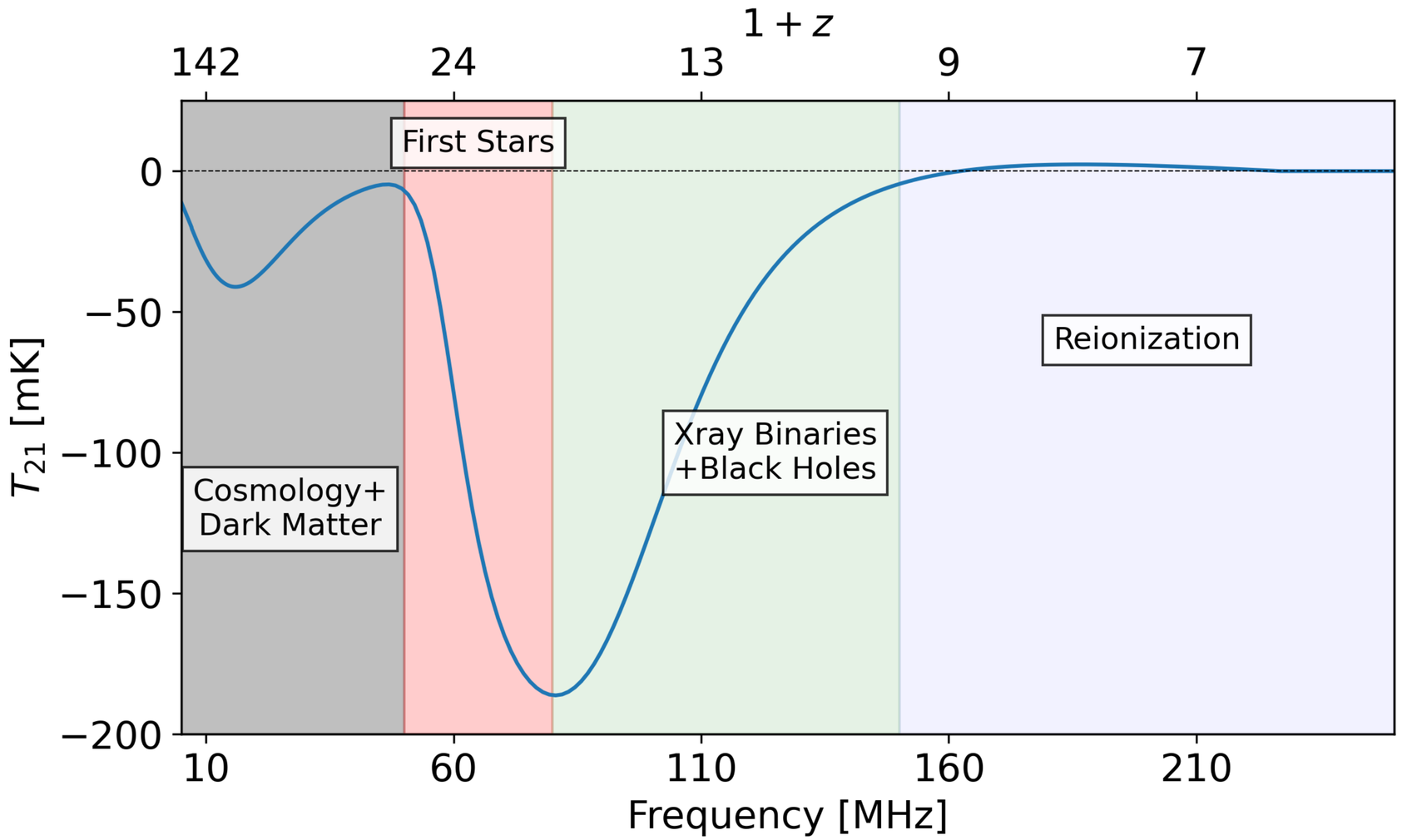
Simulating the 21-cm signal

- Several classes of simulation
  - Analytic models (Zeus21, Echo21…)
    - Make approximation
    - Evaluate in seconds
  - Semi-numerical models (21cmSPACE, 21cmFAST…)
    - Detailed grid model populated with halos and galaxies
    - Evaluate in hours
  - Hydrodynamical codes (C2-RAY…)
    - Radiative transfer codes with hydrodynamics and feedback
    - Evaluate in days to weeks



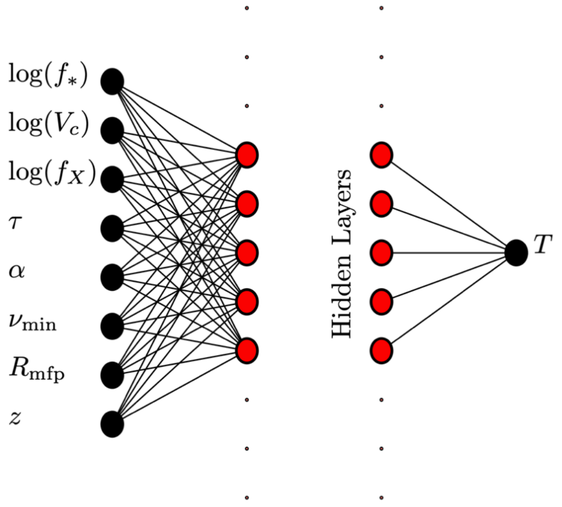


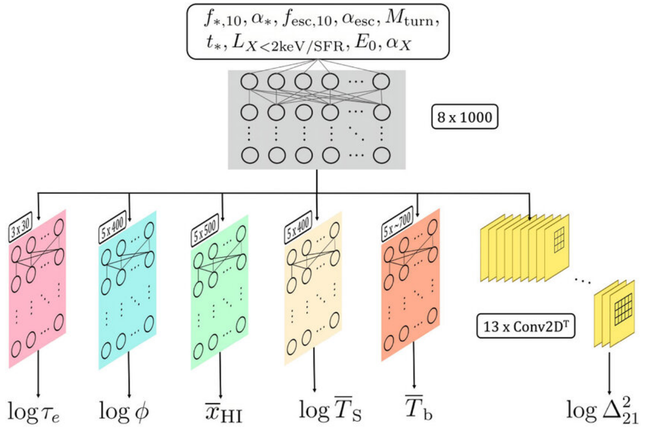
The 21-cm signal

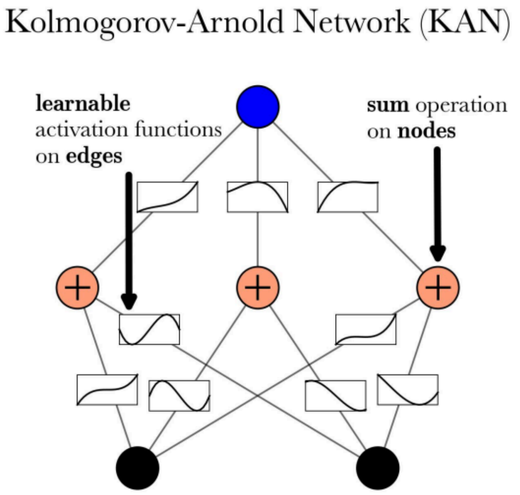


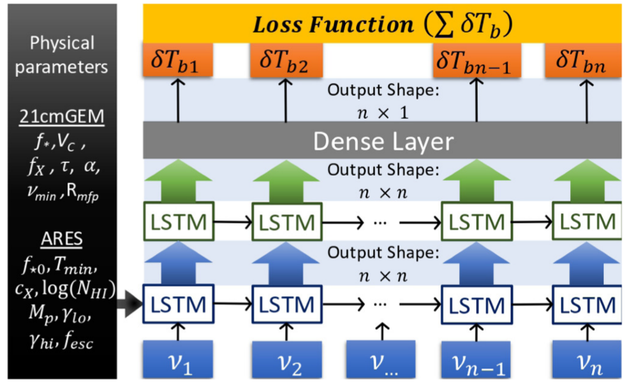
Emulators in 21cm Cosmology

- Detailed semi-numerical models take hours to evaluate
- Call them 100,000 to millions of times in inference
- Learn approximations to the model with neural networks that evaluates in milliseconds
- globalemu [Bevins+ 2021]
- 21cmVAE [Bye+ 2021]
- 21cmEMU [Breitman+ 2023]
- 21cmLSTM [Dorigo Jones+2024]
- 21cmKAN [Dorigo Jones+2025]

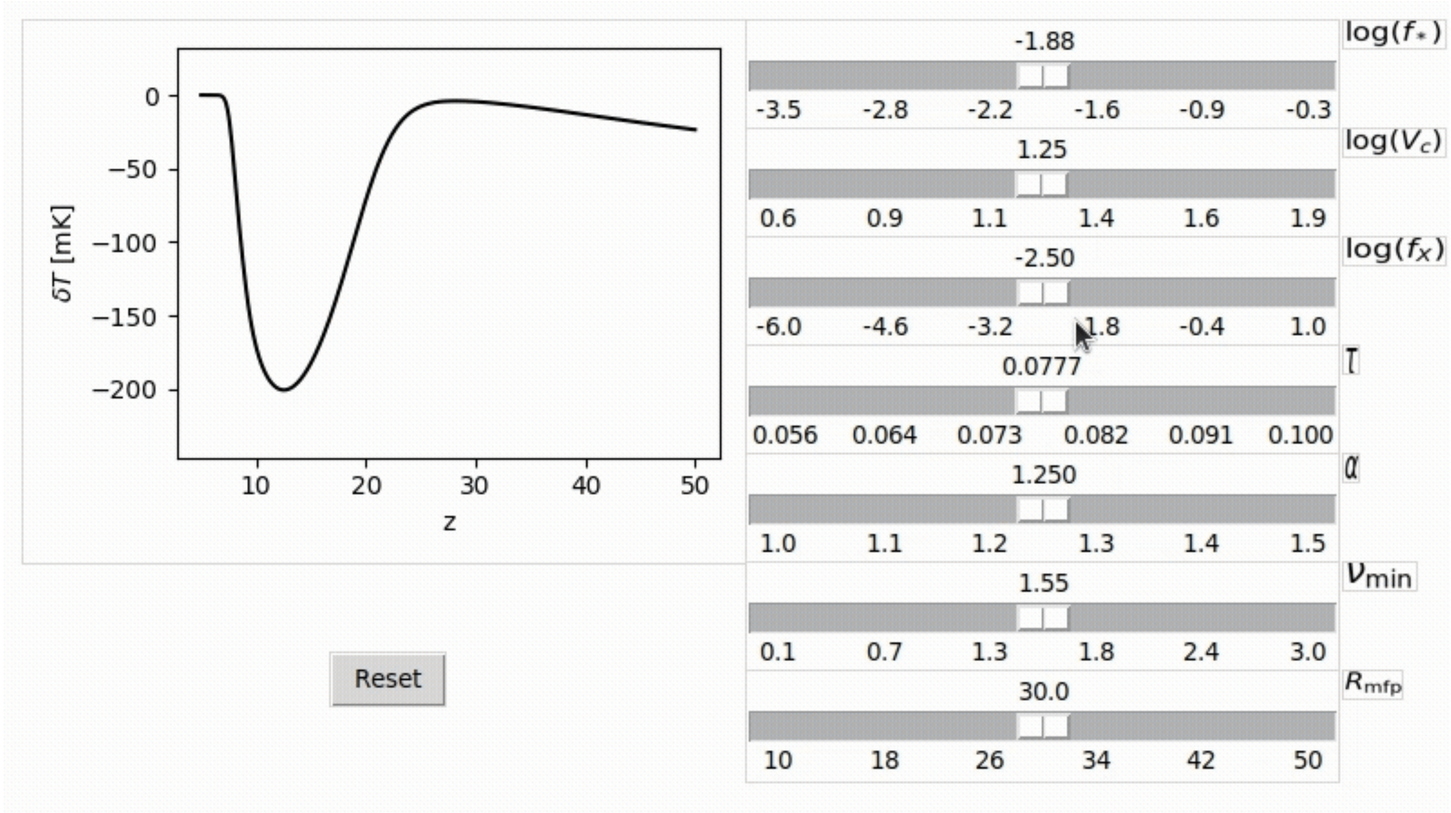








Emulators in 21cm Cosmology



Why is accuracy important?

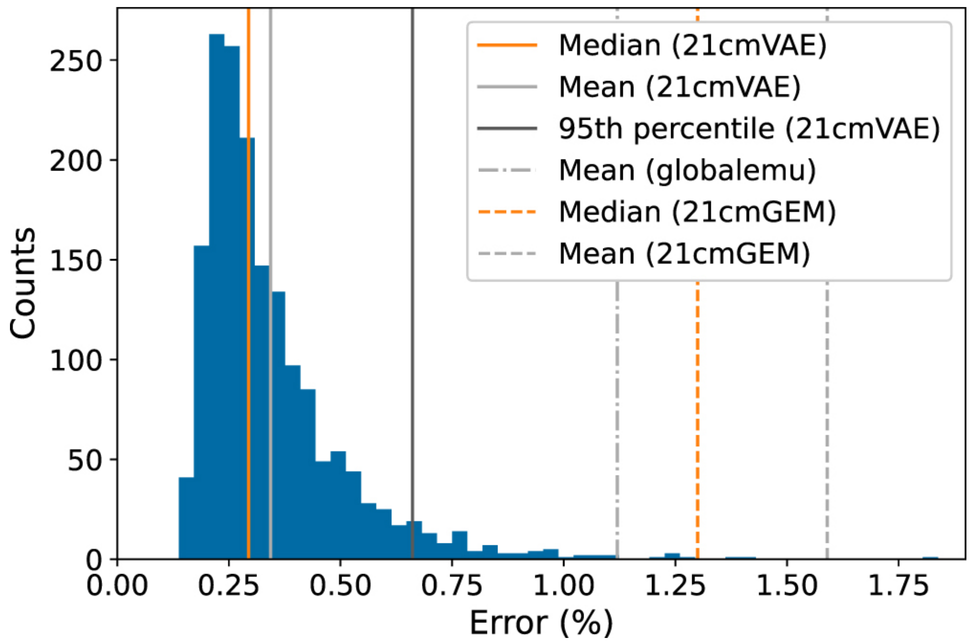
Defining required accuracy

- Typically we define accuracy with something like RMSE and a test data set

$$\epsilon = \sqrt{\frac{1}{N_\nu} \sum_i^{N_t} (T^{21}_{\rm true}(z) - T^{21}_{\rm pred}(t))^2}$$

- But what average value of ϵ over the test data is good enough for inference?

- Generally we work with “rules of thumb” e.g. globalemu paper suggested $\bar{\epsilon} \lesssim 0.1\sigma$



Why is accuracy important?

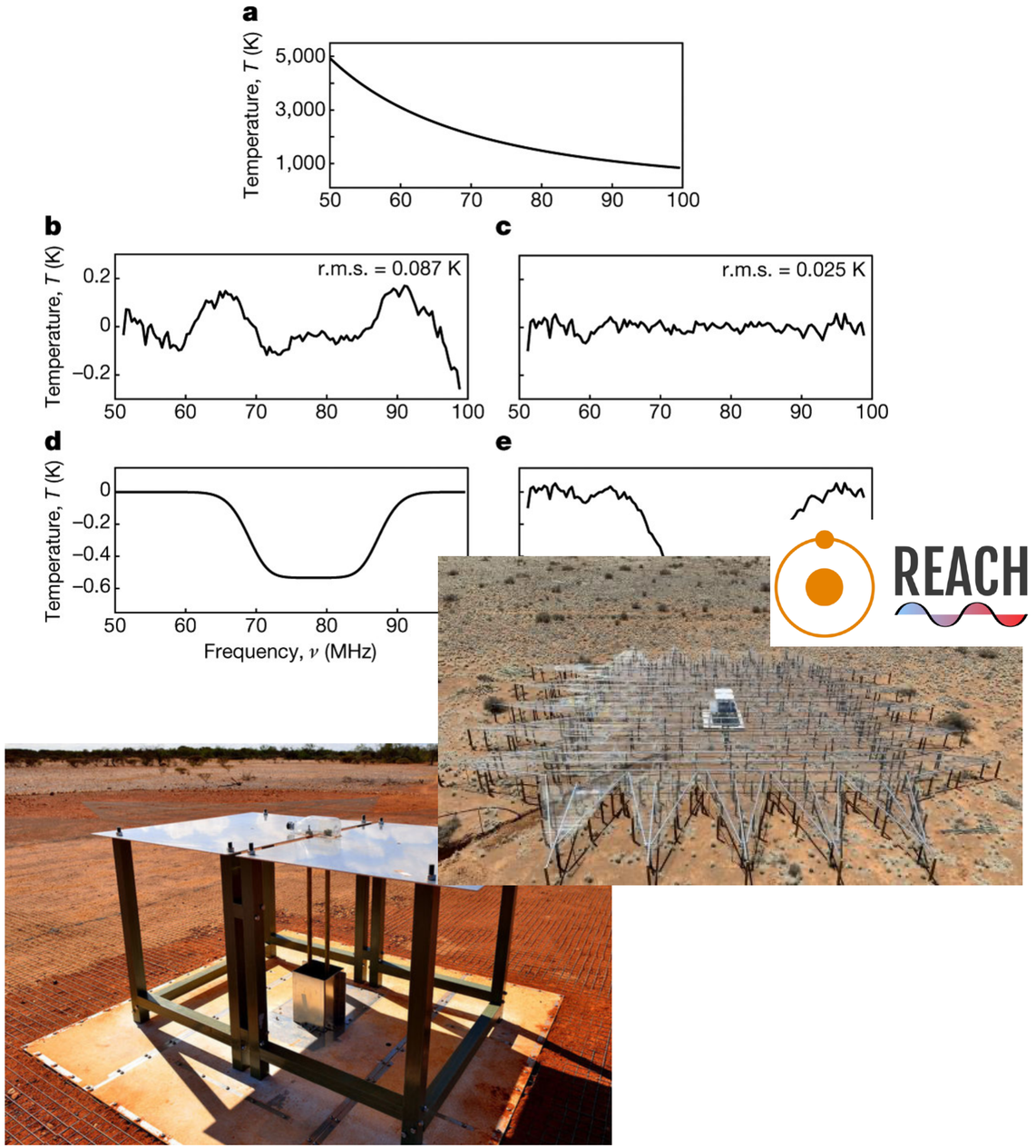
- We expect that we need to go down to around 25 mK noise to confidently detect the 21cm signal

- Most emulators have $\bar{\epsilon} \approx 1\ \mathrm{mK} \approx 0.05 \times 25\mathrm{mK}$ and it seems challenging to go beyond this

- If we assume a Gaussian likelihood and

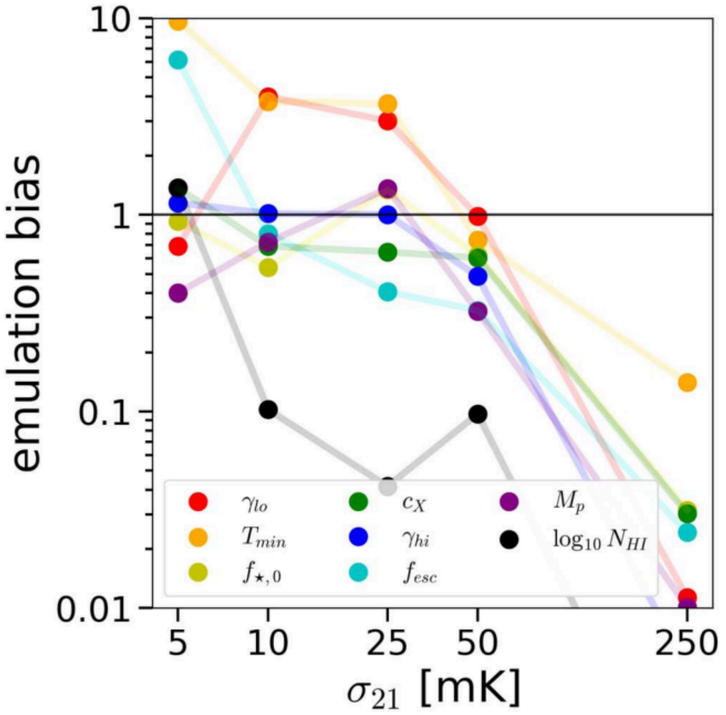
$$\sigma^2 = \sigma^2_{\mathrm{instrument}} + \bar{\epsilon}^2$$

we would expect the uncertainty from the instrument to dominate the posteriors but we really want a way to check this…



Defining required accuracy

- Dorigo Jones et al. 2023 started to ask and answer this question
- Making a direct comparison between $P(\theta \mid D, M) \leftrightarrow P(\theta \mid D, M_E)$
- We wanted to come up with something more predictive because we don't have $P(\theta \mid D, M)$ only $P(\theta \mid D, M_E)$
- "Given this error in our emulator and in our data how accurate do we expect our posteriors to be?"
- Or conversely "we want our posteriors to be this accurate and we have this level of noise in our data so how accurate does our emulator need to be?"



Impact on posterior recovery?

$$P(\theta \mid D, M) = \frac{P(D \mid \theta, M)}{P(D \mid M)} P(\theta \mid M)$$

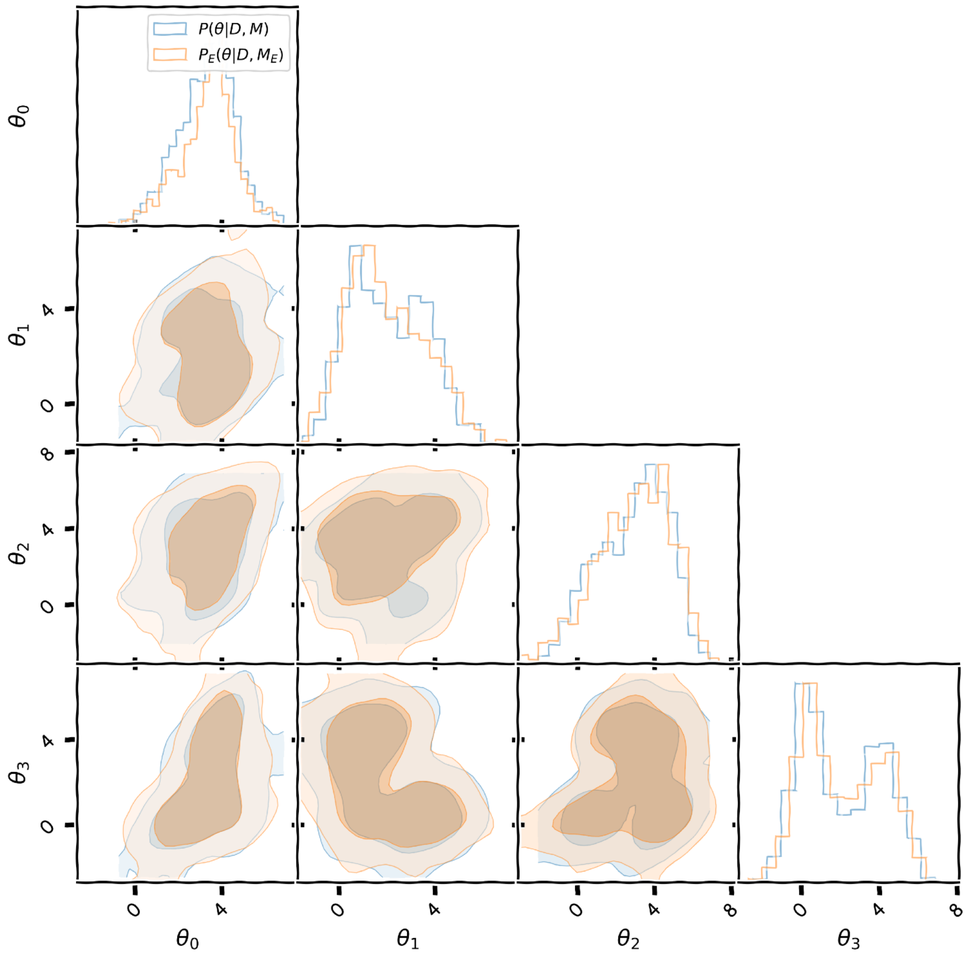
$$P = \frac{L}{Z}\pi$$

- Likelihood function is probability of the data given the model $L = P(D \mid \theta, M)$

$$\log L(M) \rightarrow \log L(M_E) + \delta \log L(M_E)$$

$$P(\theta \mid D, M) = \frac{L\pi}{\int L\pi d\theta} \rightarrow P_E(\theta \mid D, M_E) = \frac{L\pi e^{\delta \log L}}{\int L\pi e^{\delta \log L} d\theta}$$

- Is $\bar{\epsilon} \approx 0.1\sigma$ good enough?

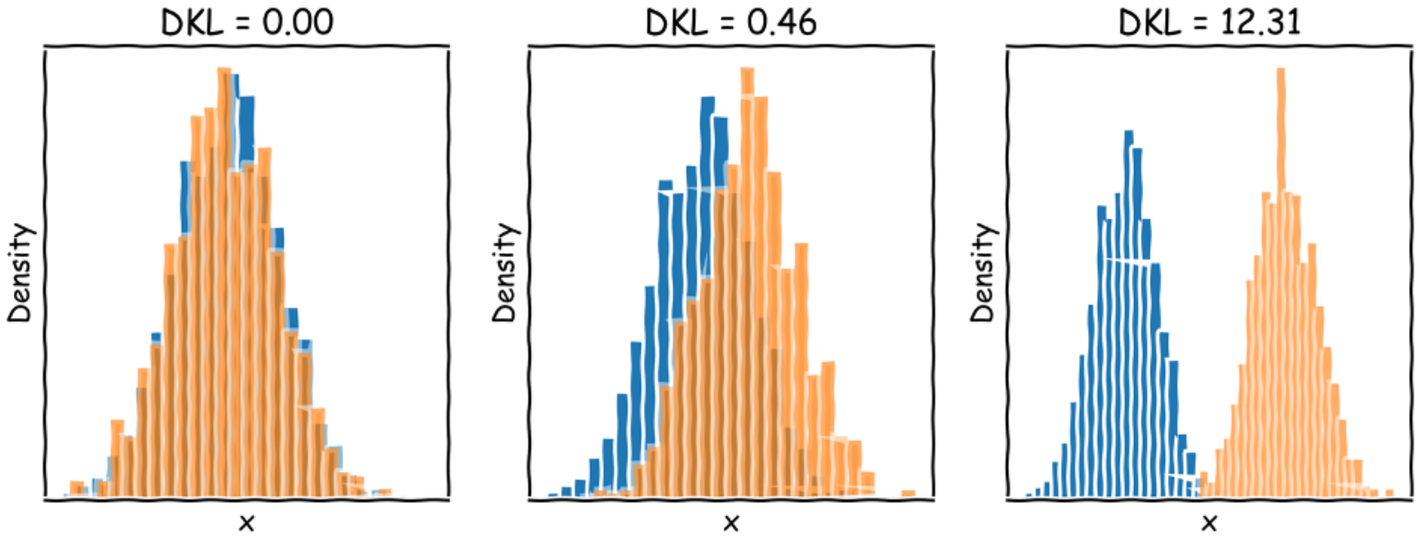


Measuring the impact of the emulator

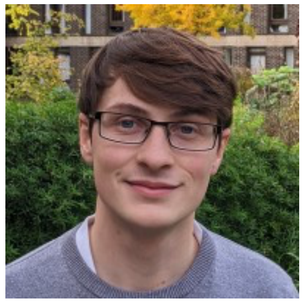
- Comprehensive measure of the difference between the true and emulated posteriors is the Kullback-Leibler Divergence

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

- However we don't have access to P...



Measuring the impact of the emulator



Thomas Gessey-Jones

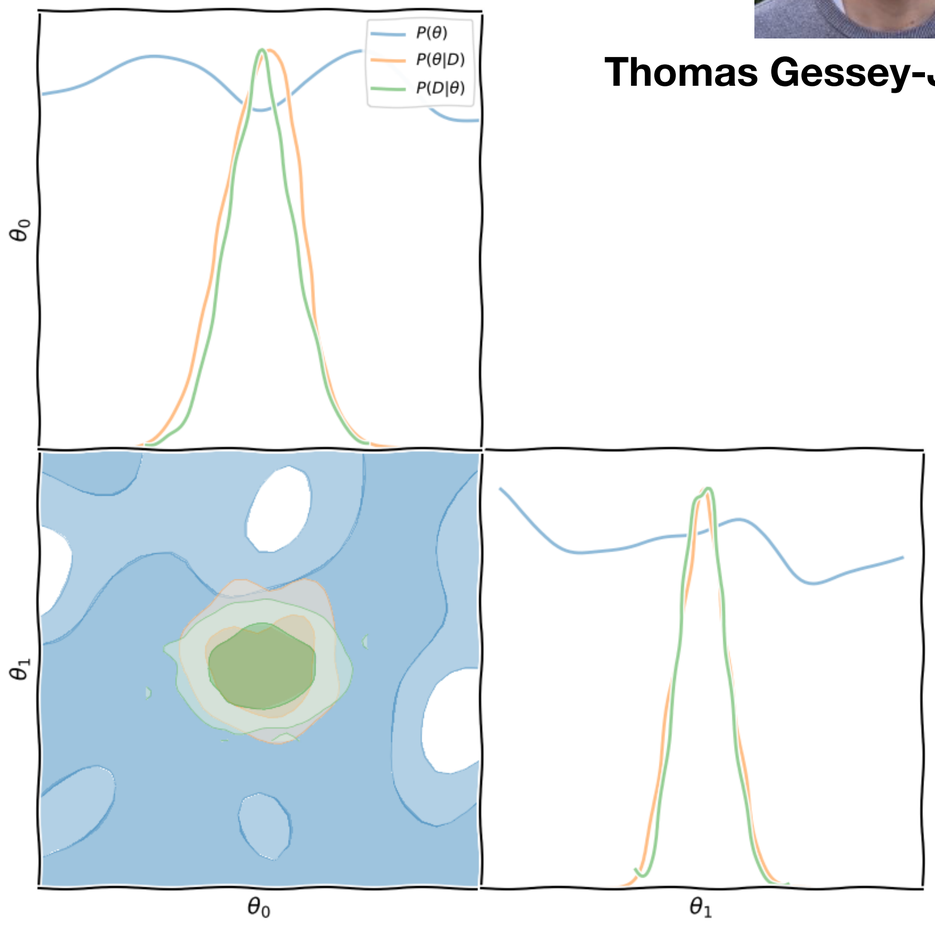
- If we make some approximations we can however define an upper limit on $D_{KL}(P||P_E)$

$$L = \mathcal{N}(D; \Sigma, M(\theta))$$
$$\pi = \mathcal{U}(\theta)$$
$$\rightarrow P = \mathcal{N}(\theta; C, \mu)$$

- P and P_E are Gaussian then the KL divergence between them is given by

$$D_{KL} = \frac{1}{2}\left[\log\left(\frac{|C_E|}{|C|}\right) - N_\theta + \mathrm{tr}(C_E^{-1}C) + (\mu_E - \mu)^T C^{-1}(\mu_E - \mu)\right]$$

- Where C, C_E, μ and μ_E are functions of M, M_E and Σ the noise in the data and $M_E(\theta) = M(\theta) + E(\theta)$



Measuring the impact of the emulator

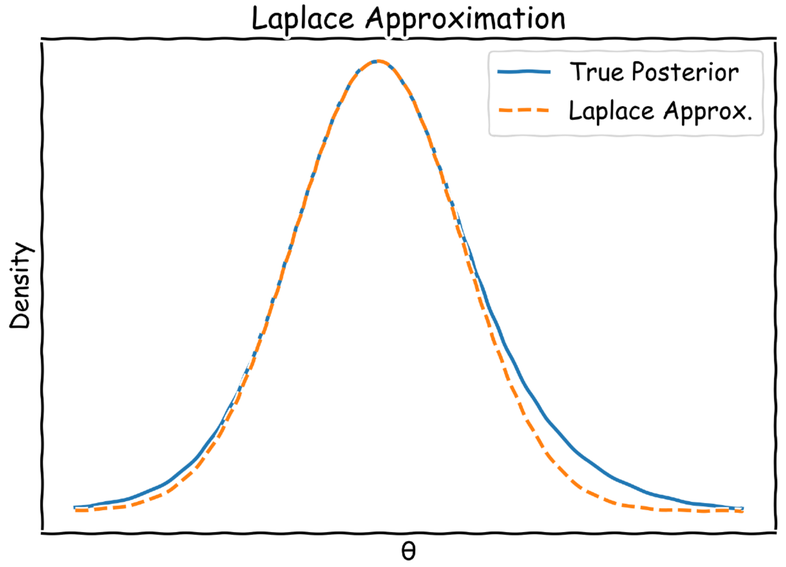
- Assume a linear model and linear emulator error

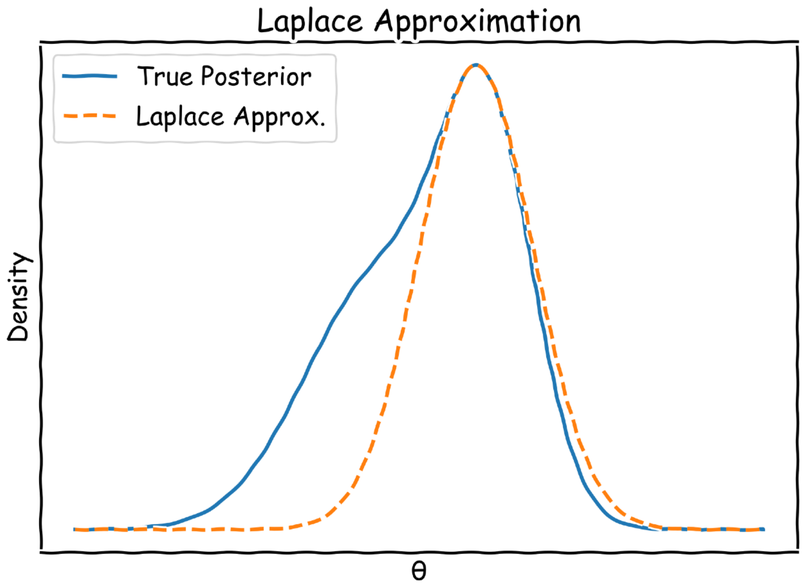
$$\mathscr{M}(\theta) \approx M\theta + m \text{ and } E(\theta) \approx E\theta + \epsilon$$

Such that $M_\epsilon(\theta) = (M + E)\theta + (m + \epsilon)$

- Comes from Taylor expansion of model around the MAP and the assumption that the posterior is sharply peaked so we can ignore higher order terms

$$M = \mathscr{J}(\theta_0)$$
$$m = M(\theta_0) - \mathscr{J}(\theta_0)\theta_0$$





Measuring the impact of the emulator

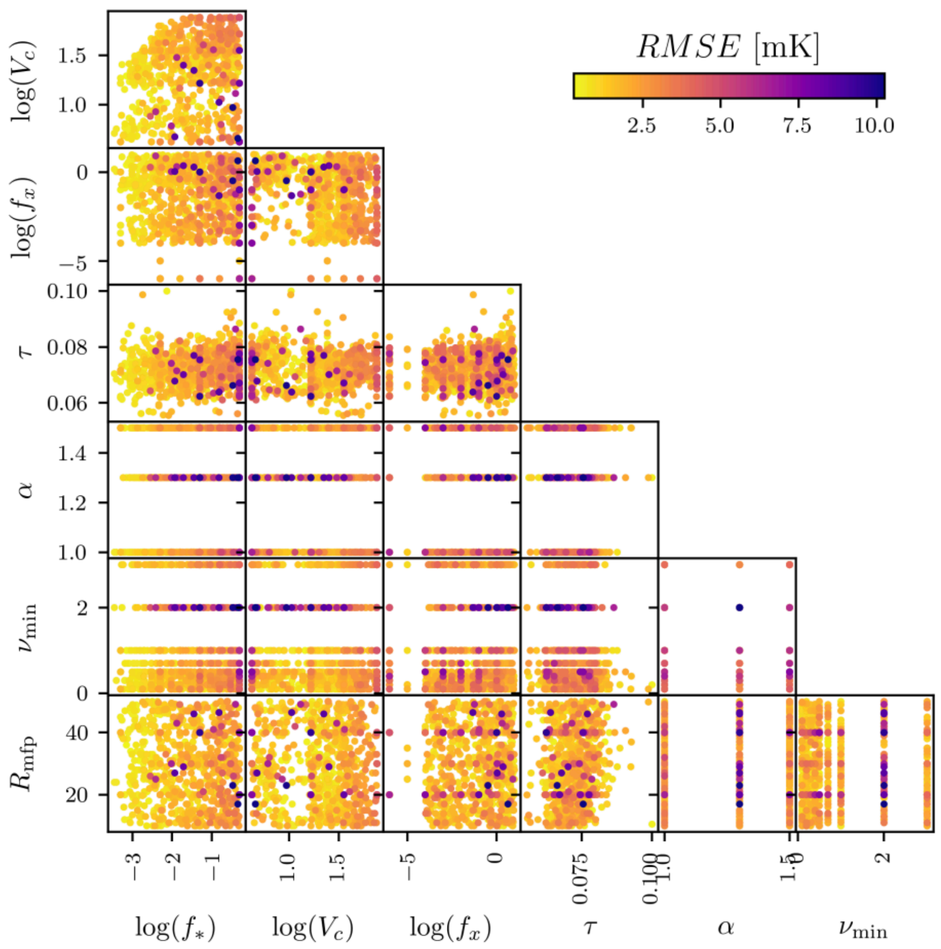
$$\mathscr{M}(\theta) \approx M\theta + m \text{ and } E(\theta) \approx E\theta + \epsilon$$

Such that $M_\epsilon(\theta) = (M + E)\theta + (m + \epsilon)$

- Then assume that the emulator error evolves slowly over the parameters space relative to the model then $E(\theta) \approx \epsilon$
- Substitute into the posteriors

$$P \propto \exp\left(-\frac{1}{2}(D - m - M\theta)^T \Sigma^{-1} (D - m - M\theta) \right)$$

Then we can express C, C_E, μ and μ_E in terms of M, m and ϵ.



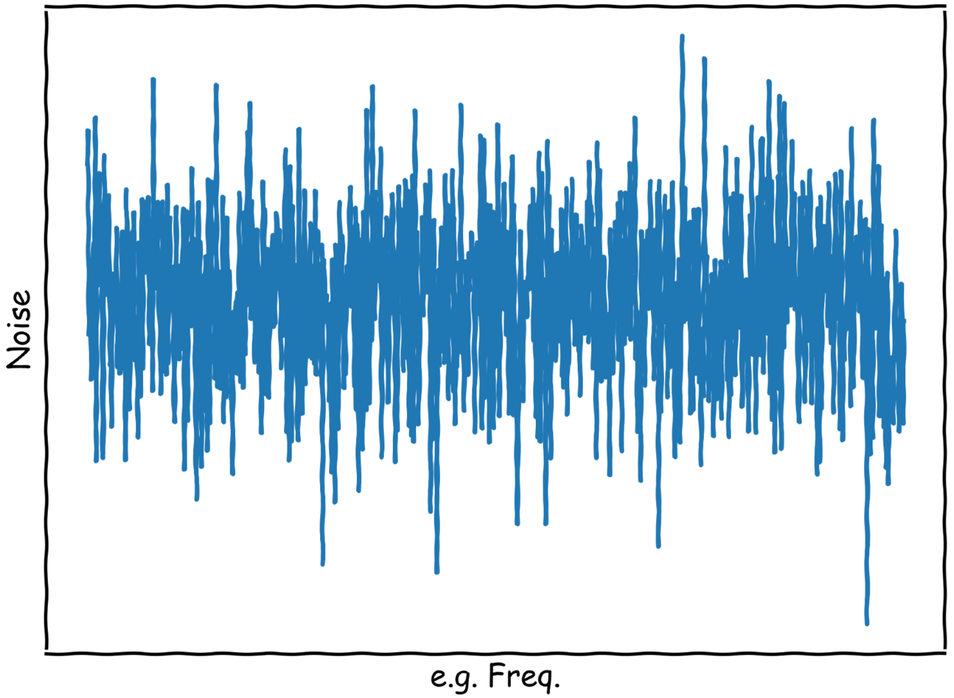
Measuring the impact of the emulator

- We then assume the noise $\Sigma = \frac{1}{\sigma^2}\mathbf{1}_{N_d}$
- And substitute everything into the expression for D_{KL}

$$D_{KL} = \frac{1}{2}\frac{1}{\sigma^2}\epsilon^T MM^+\epsilon$$

- From here we can rotate into the orthogonal eigenbasis of MM^+ and you end up with an inner product over a diagonal of a matrix of positive values

$$D_{KL} = \frac{1}{2}\frac{1}{\sigma^2}\sum_i U^{-1}\epsilon\lambda_i$$



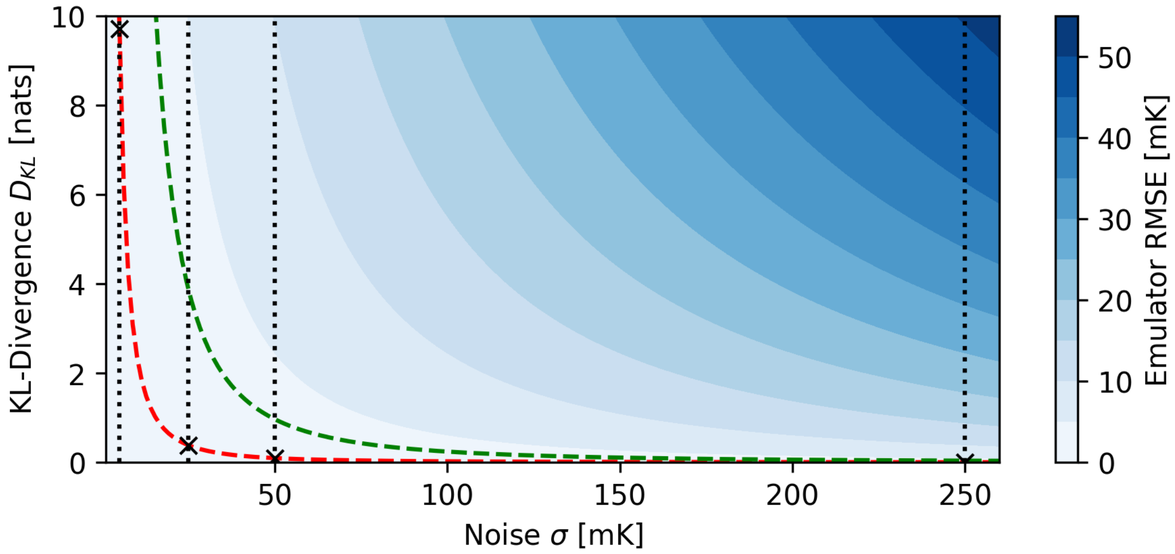
Measuring the impact of the emulator

$$D_{KL} = \frac{1}{2}\frac{1}{\sigma^2}\sum_i (U^{-1}\epsilon)^2 \lambda_i$$

- We can then say $D_{KL} \leq \frac{1}{2}\frac{1}{\sigma^2}\max(\lambda_n)\,||\,U^{-1}\epsilon\,||^2$
- U^{-1} leaves $||\epsilon||$ unchanged and MM^+ is a projection matrix so its eigenvalues are either 1 or 0

$$D_{\mathrm{KL}}(P\,||\,P_E) \leq \frac{1}{2}\frac{1}{\sigma^2}\,||\epsilon||^2$$

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

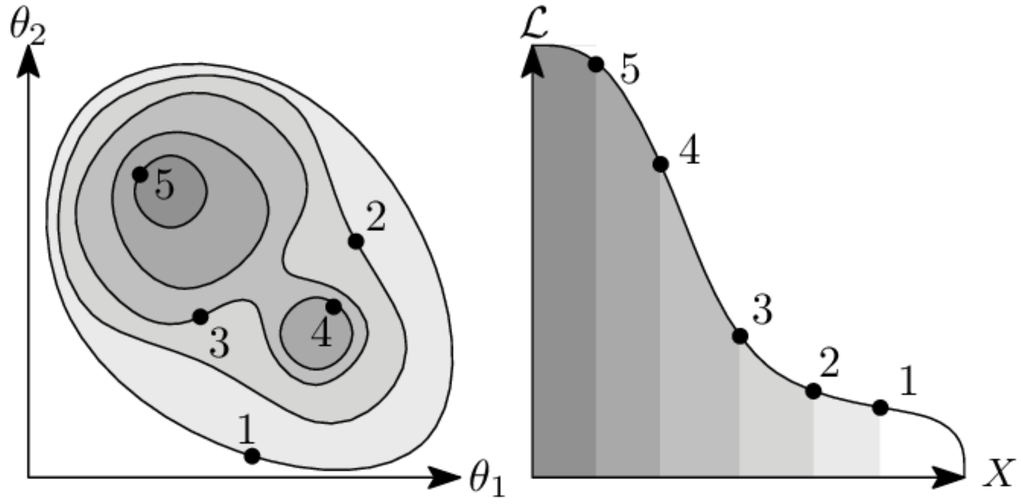


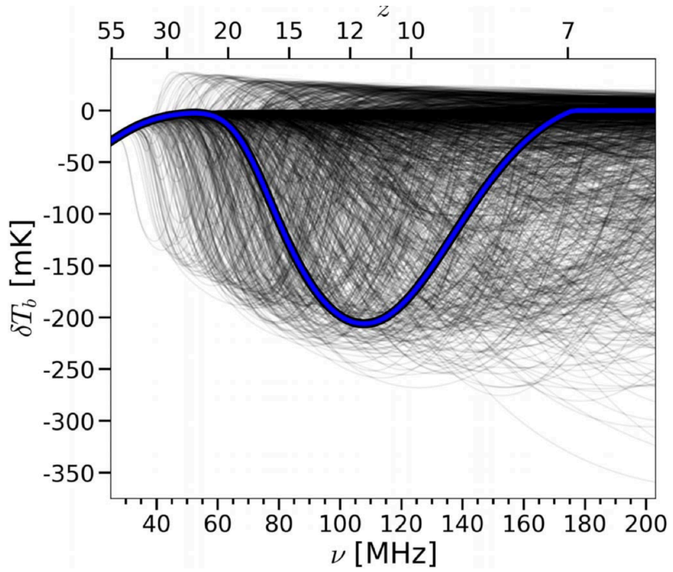
Measuring the impact of the emulator

$$\epsilon = \sqrt{\frac{1}{N_\nu} \sum_i^{N_t} (S_{\text{true}}(t) - S_{\text{pred}}(t))^2} \longrightarrow D_{\text{KL}}(P \,||\, P_E) \leq \frac{N_d}{2} \left(\frac{\text{RMSE}}{\sigma} \right)^2$$

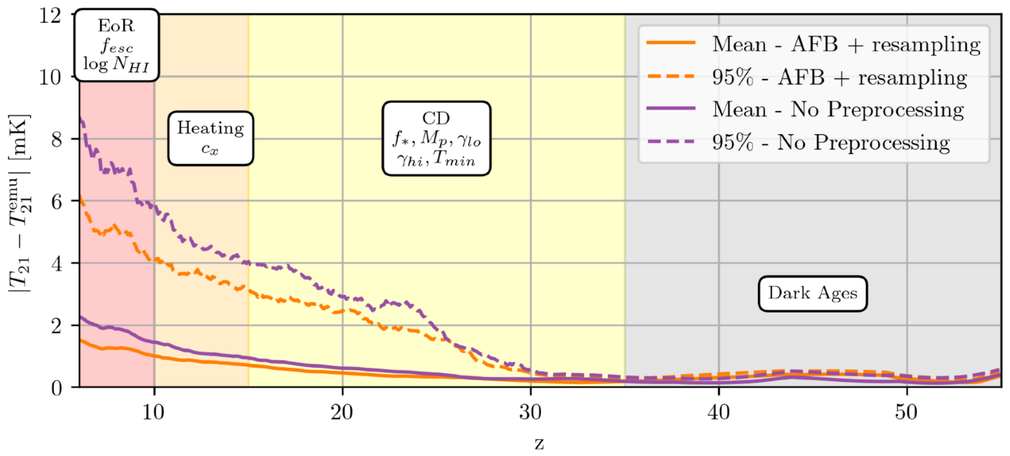
Testing on a 21cm Cosmology problem

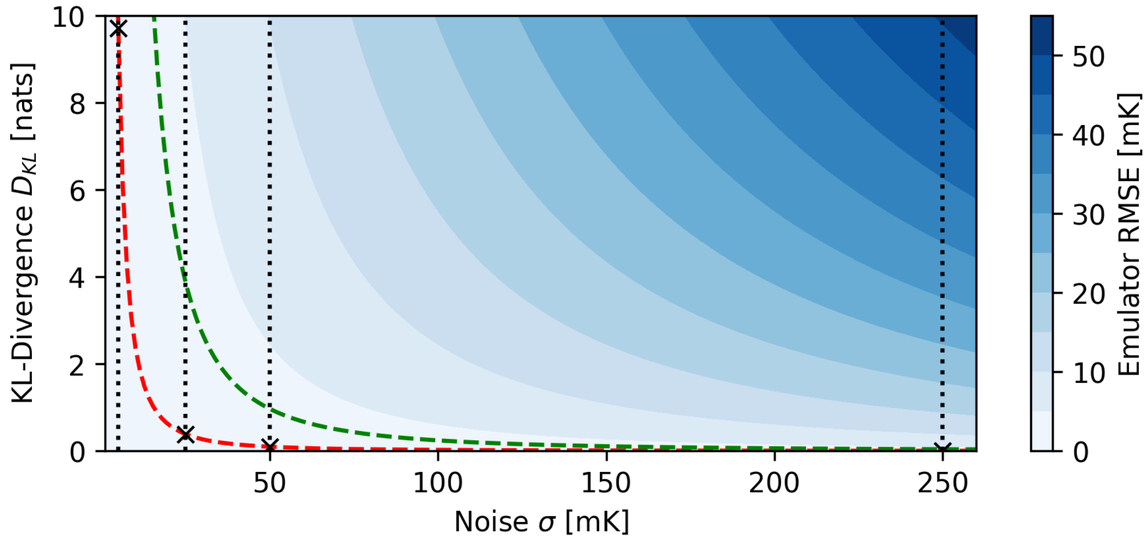
- Assuming the data comprises of signal plus noise
- Using the ARES 1D radiative transfer code with 8 parameters
- Using Polychord to perform inference with a gaussian likelihood
- Assuming absolute knowledge of the level of noise in the data
- Running for 5, 25, 50 and 250 mK noise





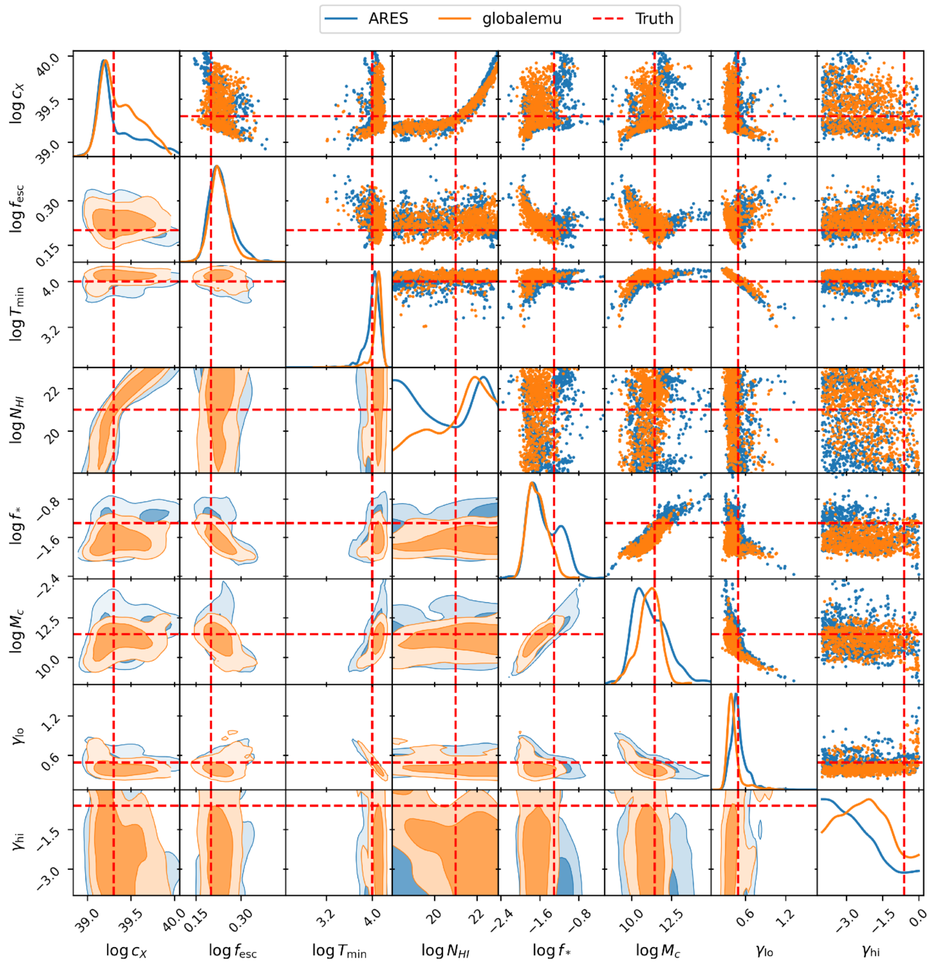
globalemu performance and ARES modelling

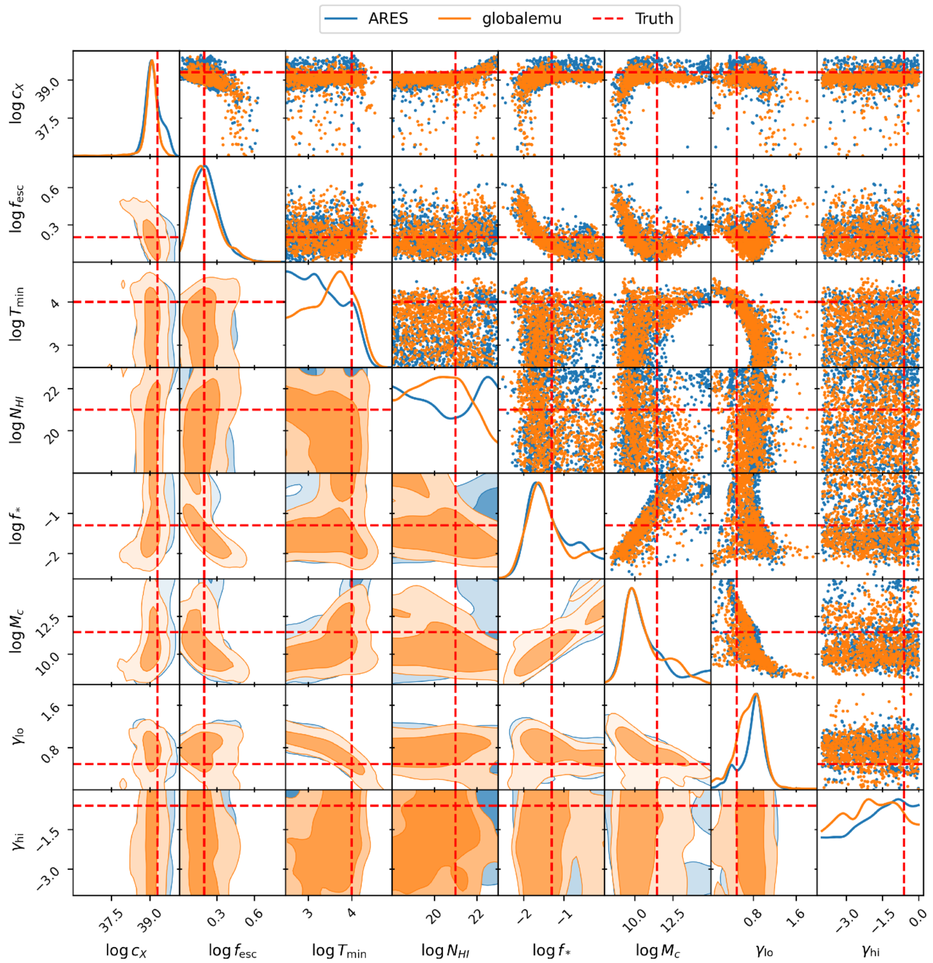




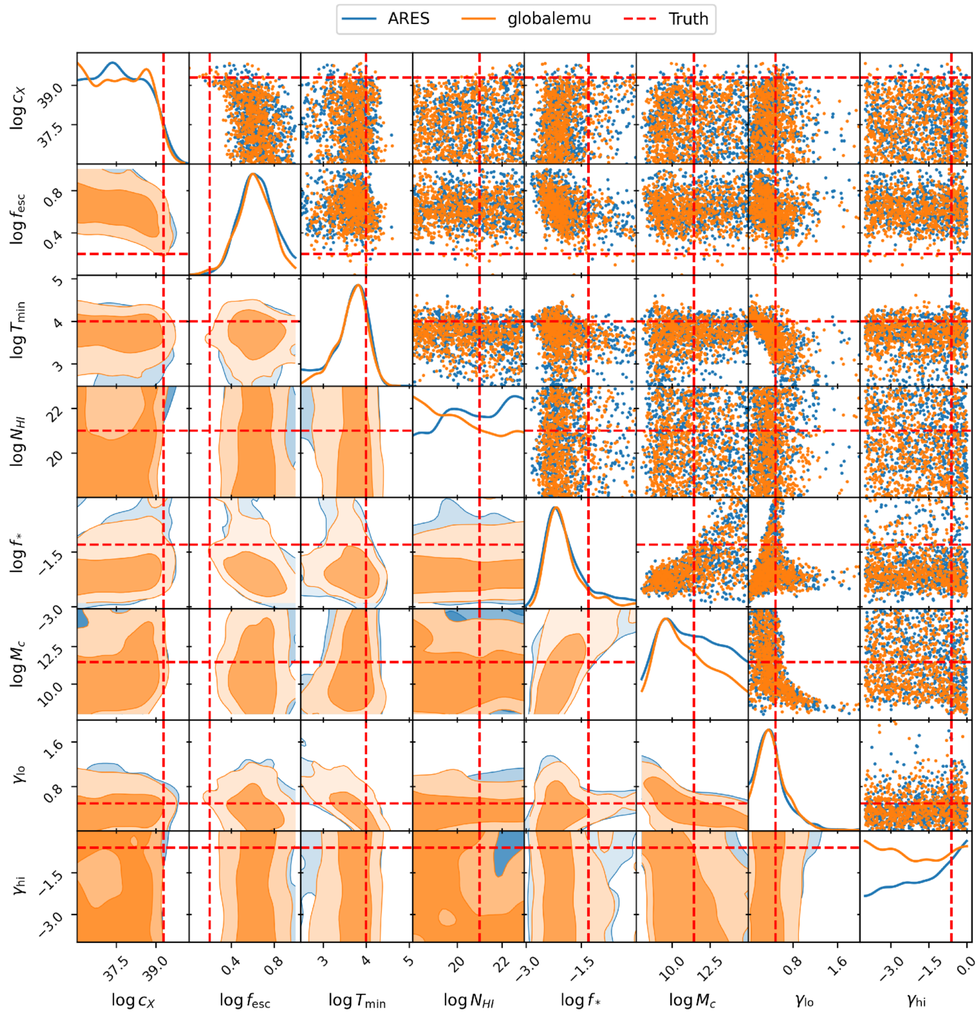
Noise Level [mK]	Estimated $\mathcal{D}_{\mathrm{KL}} \leq$	
	Mean RMSE	95th Percentile
5	9.60	96.62
25	0.38	3.86
50	0.10	0.97
250	0.004	0.039

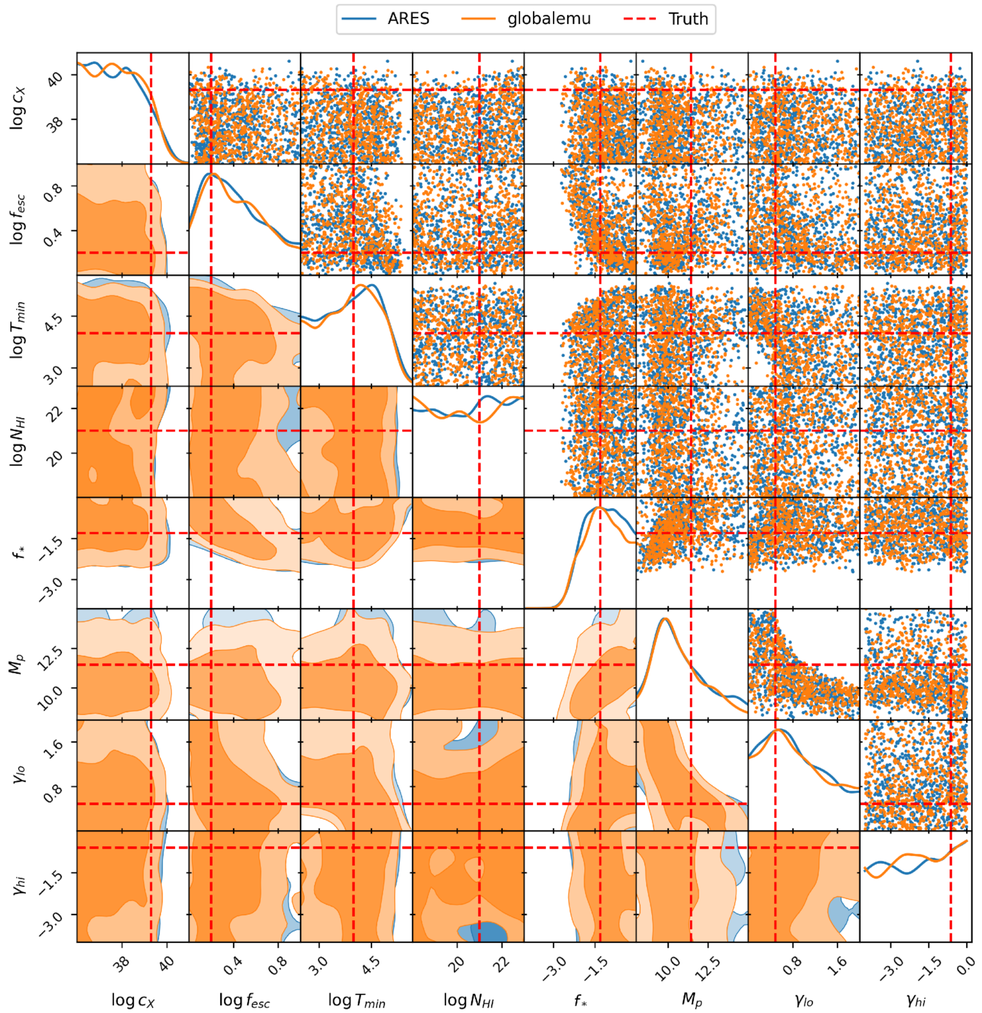
Running the analysis - 5 mK and 25 mK





Running the analysis - 50 and 250 mK

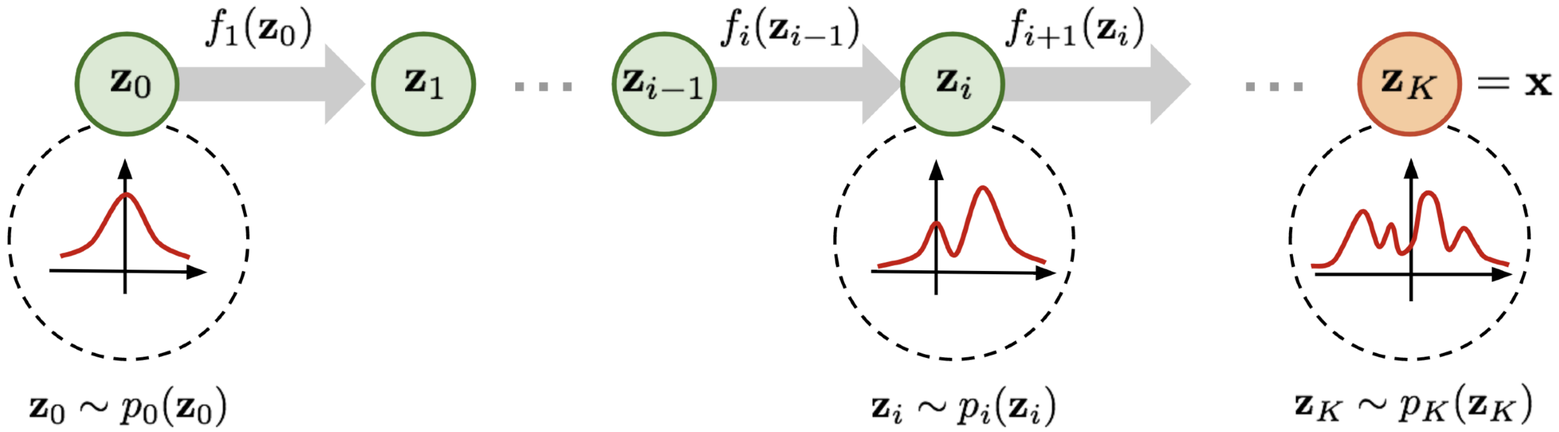




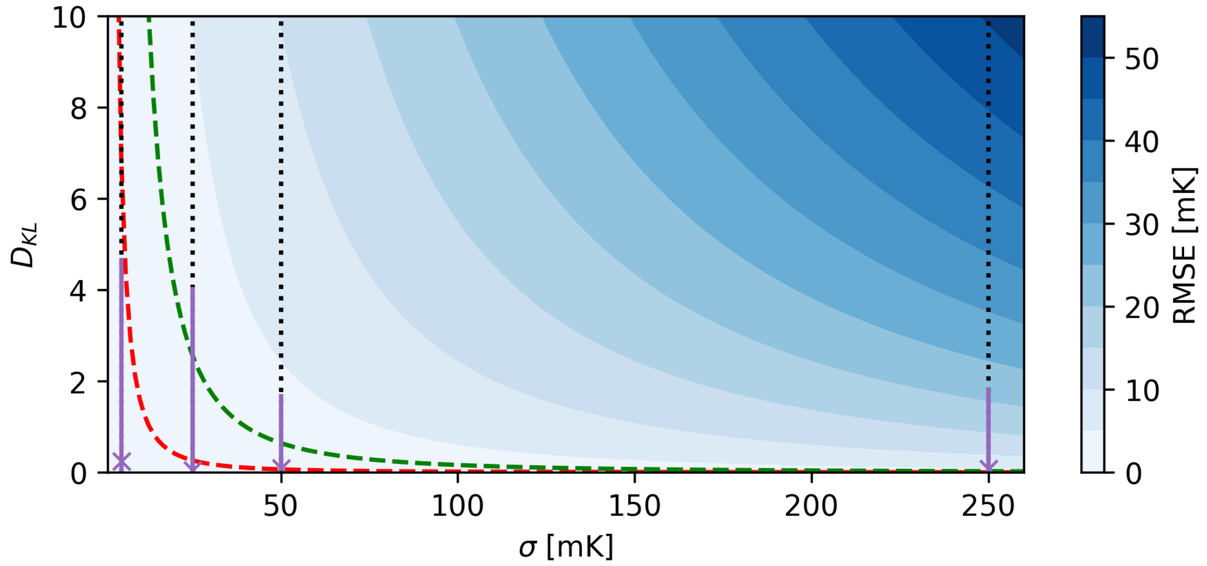
arXiv:2503.13263

Measuring the D_{KL} with normalising flows

- Need to be able to evaluate the probability on each distribution for the same samples
- Use normalising flows implemented with ***margarine*** [see Bevins et al 2022, 2023, arXiv:2207.11457, arXiv:2205.12841]



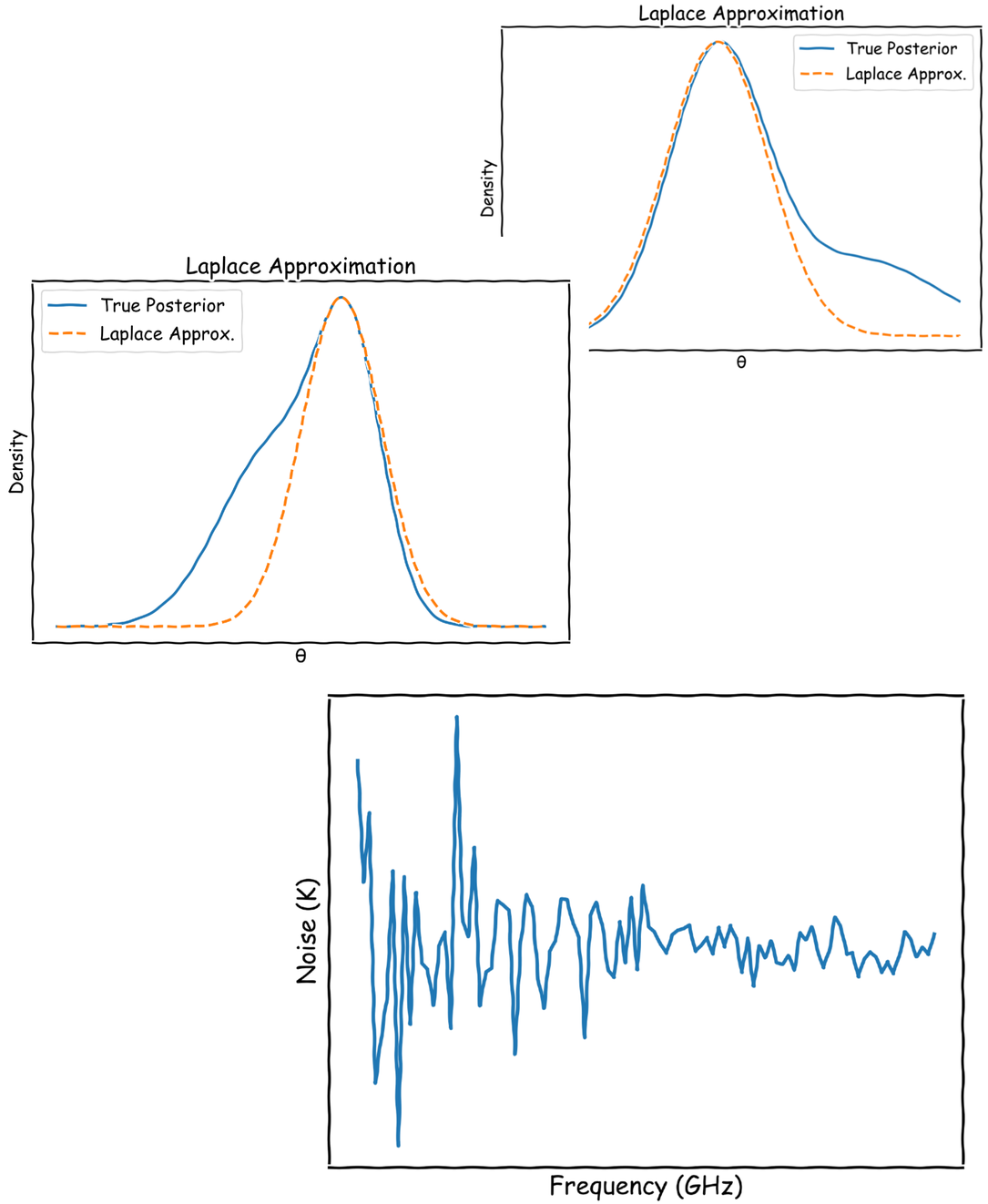
How about the D_{KL}?



Noise Level [mK]	Estimated $\mathcal{D}_{\mathrm{KL}} \leq$		Actual $\mathcal{D}_{\mathrm{KL}}$
	Mean RMSE	95th Percentile	
5	9.60	96.62	$0.25^{+4.45}_{-0.25}$
25	0.38	3.86	$0.05^{+4.02}_{-0.52}$
50	0.10	0.97	$0.09^{+1.62}_{-0.03}$
250	0.004	0.039	$0.08^{+1.78}_{-0.02}$

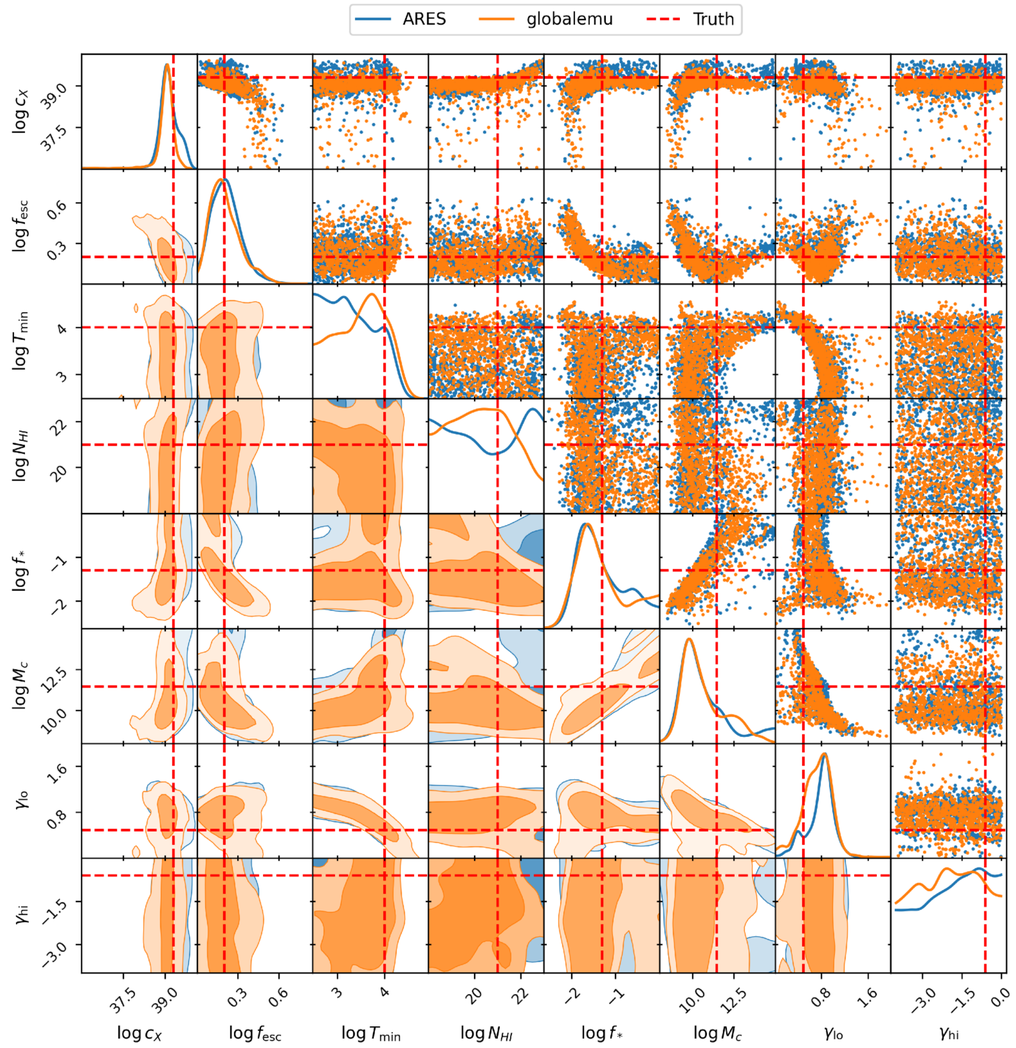
Limitations of the approximation

- The approximation assumes linearity around the peak of the posterior which might not hold in higher dimensions
- Posteriors become curved or multi modal
- Assuming a Gaussian likelihood and posterior
- Assumes uncorrelated noise in the data
- Assumes noise is constant across the data

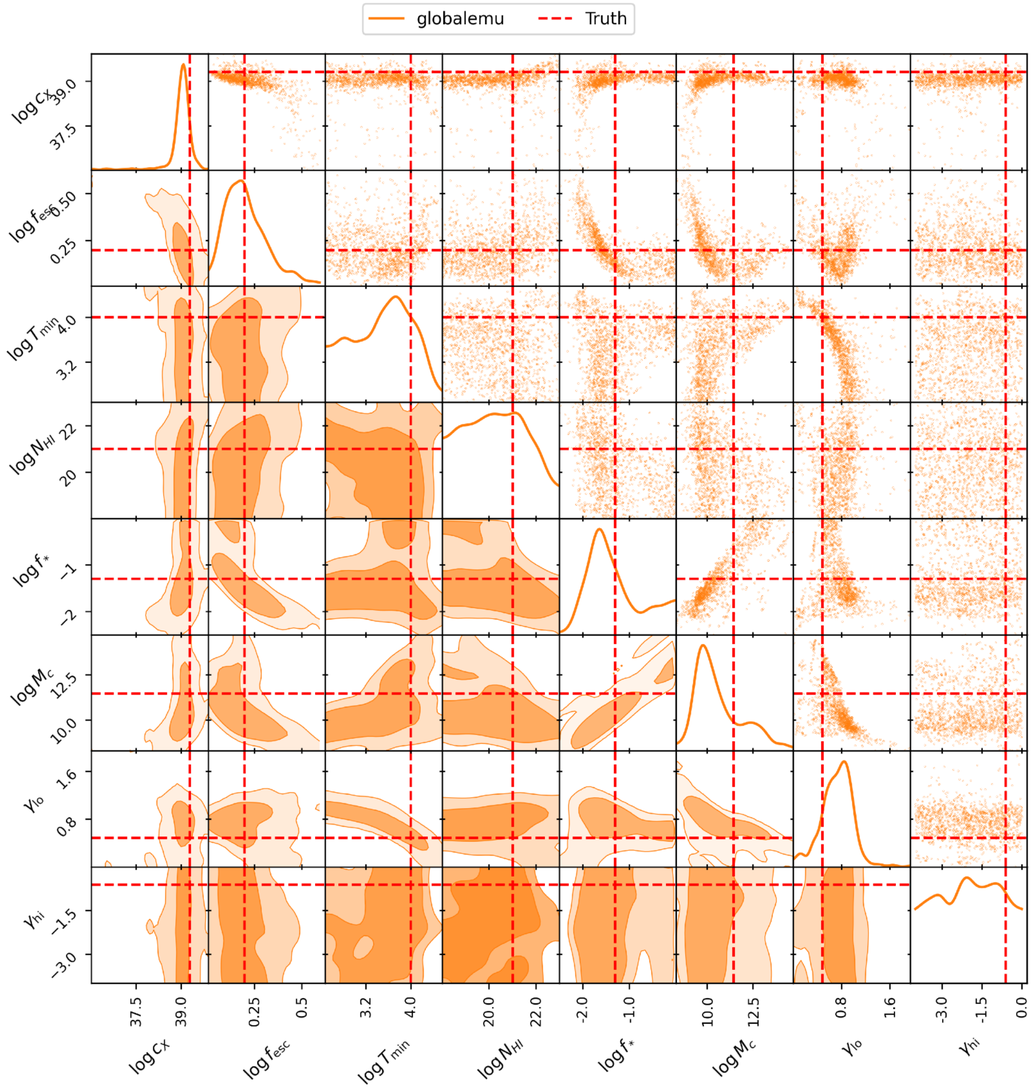


How to use this?

Build Confidence in inference



Build Confidence in inference



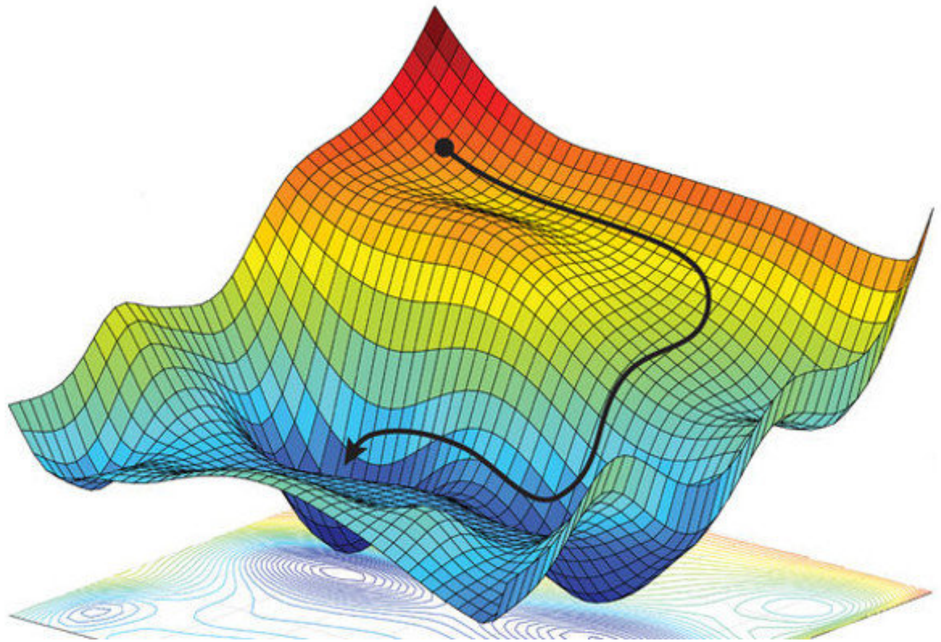
As a Loss Function and for hyperparameter tuning?

- If we know how noisy our data is we can use

$$D_{KL} \leq \frac{N_d}{2}\left(\frac{\mathrm{RMSE}}{\sigma}\right)^2$$

as a loss function in our emulator training

- Similarly we can use it as our objective in hyperparameter tuning with frameworks like optuna



Conclusions

- A useful upper bound on the incurred information loss from using emulators in inference
- We demonstrated that we can accurately recover posteriors even with $\bar{\epsilon} \approx 0.2\sigma$ for 21cm
- Broadly applicable beyond 21cm
- Can use this as a loss function or for hyperparameter tuning
- Accepted in MNRAS [arXiv:2503.13263]
- https://github.com/htjb/validating_posteriors