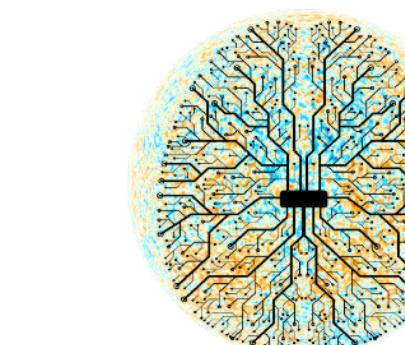


On the accuracy of posterior recovery with neural network emulators

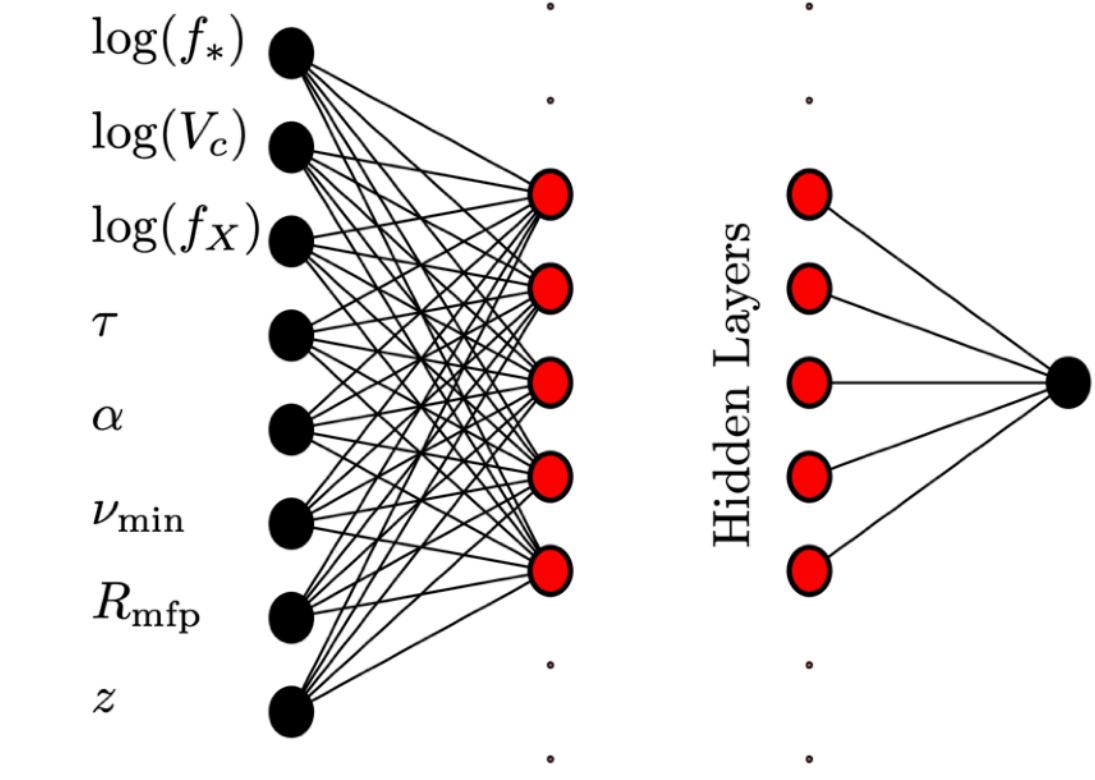
Harry Bevins

With Thomas Gessey-Jones and Will Handley

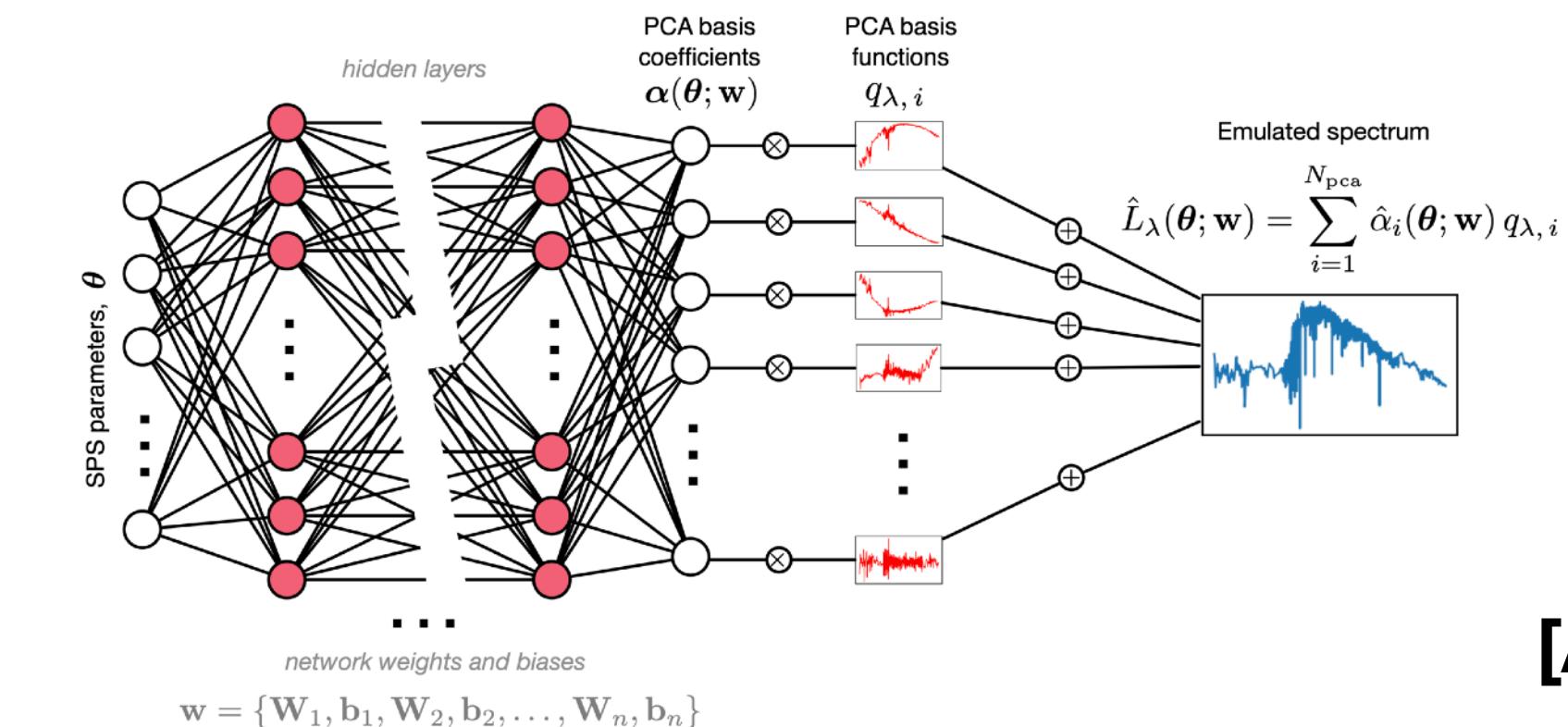


Emulators in Cosmology and Astrophysics

- Neural network emulators are really important in Cosmology and Astrophysics
- For fast inference on computationally expensive likelihoods
- Generating large training data sets for training simulation based inference algorithms



Cosmopower
[Spurio Mancini+2021]

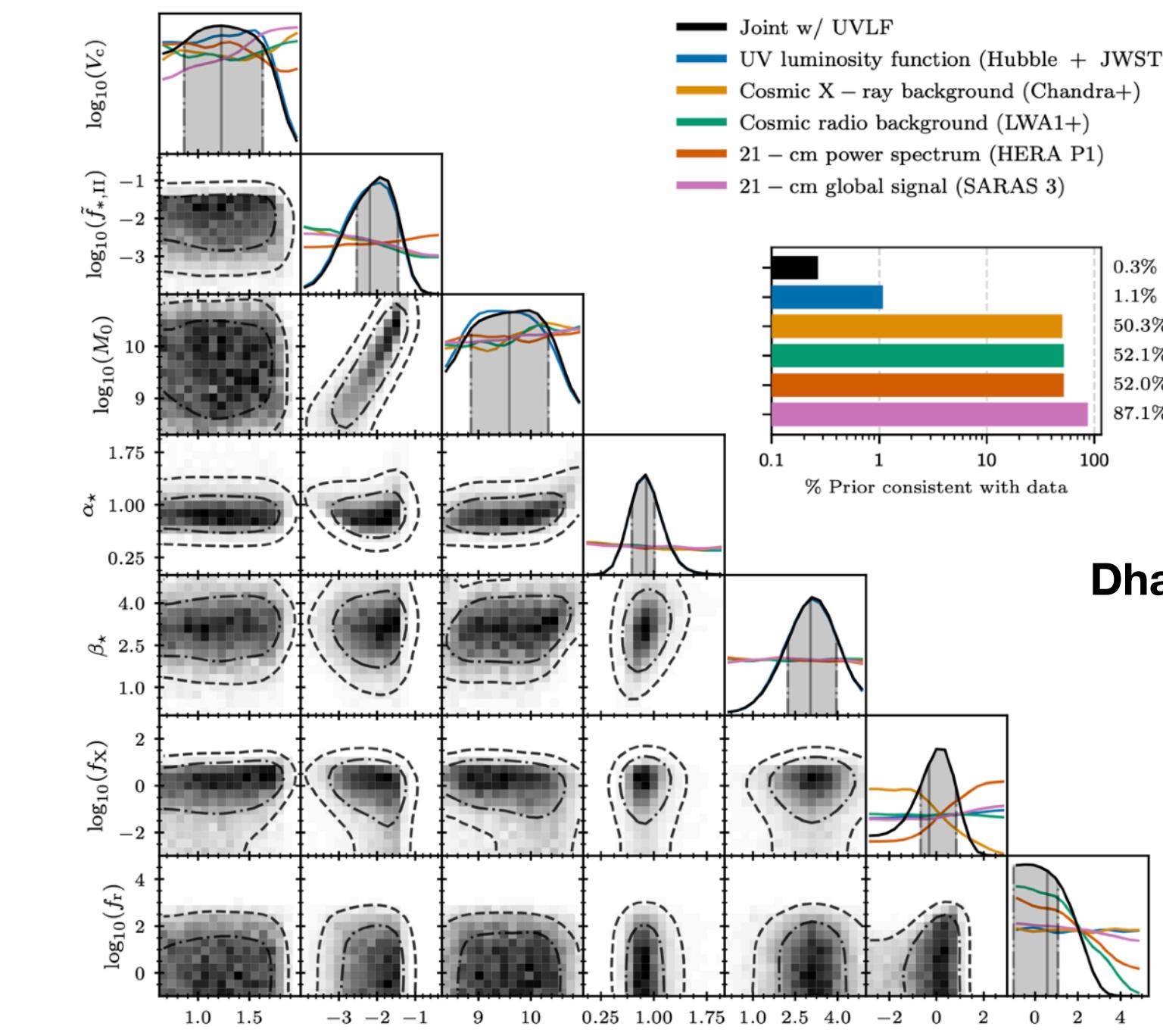


Speculator
[Alsing+2020]

globalemu [Bevins+21]
21cmLSTM [Dorigo Jones+2024]
21cmEMU [Breitman+ 2023]
21cmGEM [Cohen+2017]
And **21cmVAE** [Bye+2022]

Emulators in Cosmology and Astrophysics

- In this work we are focused on likelihood based inference
- Semi-numerical simulations of cosmological signals are very computationally expensive
- Train emulators on example simulations and use these the likelihood functions
- Established method for doing inference



Dhandha, Gessey-Jones,
Bevins et al. 2025

(a) Cosmic shear with 37 (Λ CDM) and 39 (w_0w_a CDM) parameters, described in Sect. 4.

Method	$\log(z_{\Lambda\text{CDM}})$	$\log(z_{w_0w_a\text{CDM}})$	$\log \text{BF}$	Total computation time
CAMB + nested sampling	-107.03 ± 0.27	-107.81 ± 0.74	0.78 ± 0.79	~ 8 months (48 CPUs)
CosmoPower-JAX + NUTS + harmonic	40956.55 ± 0.06	40955.03 ± 0.04	1.53 ± 0.07	2 days (sampling, 12 GPUs) + 12 minutes (evidence, 1 GPU + 48 CPUs)
CosmoPower-JAX + NUTS + naïve flow estimator	400958 ± 5	40957 ± 4	1 ± 6	Similar to harmonic

Piras et al 2024

(b) 3x(3x2pt) with 157 (Λ CDM) and 159 (w_0w_a CDM) parameters, described in Sect. 5.

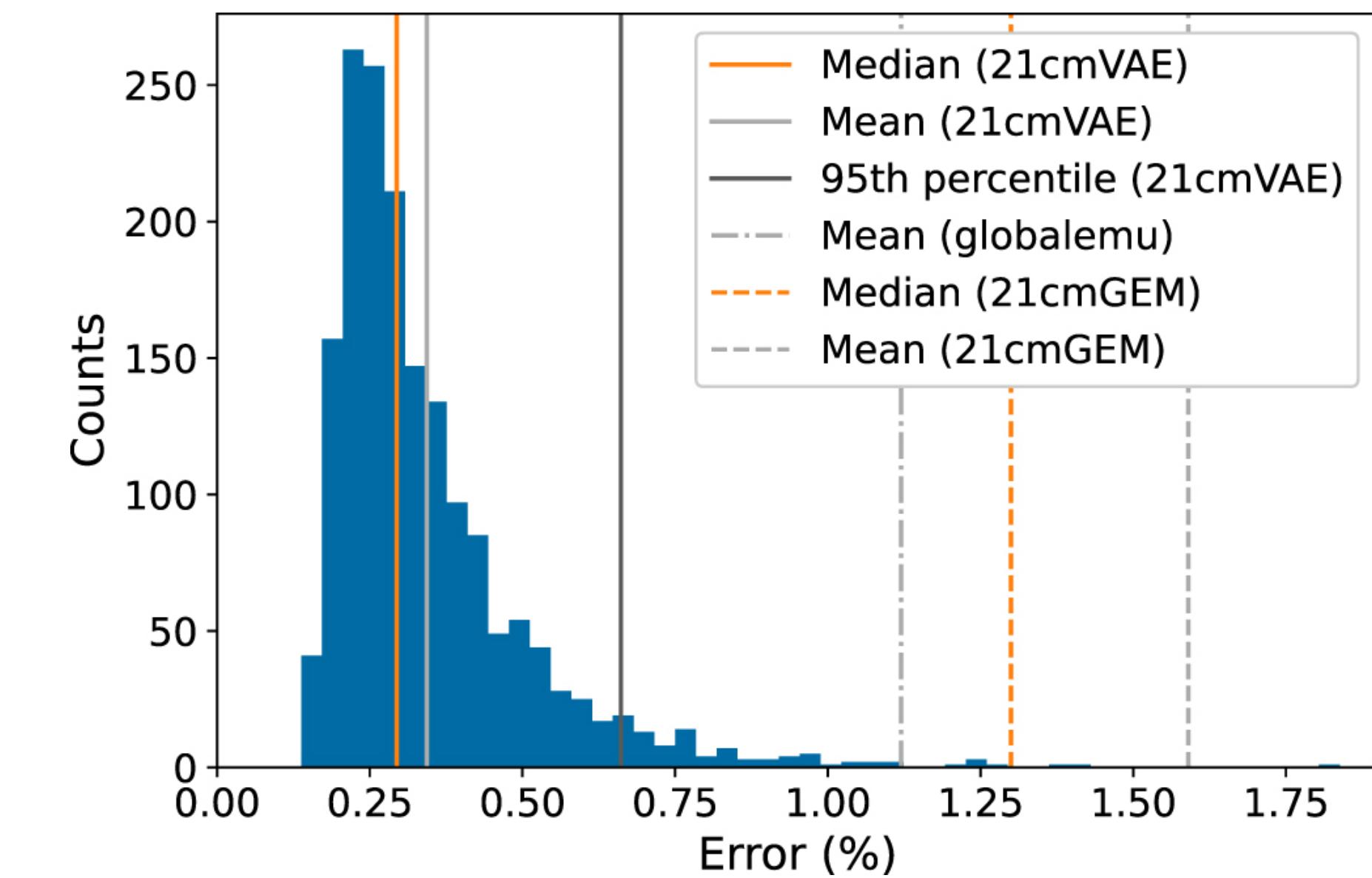
Method	$\log(z_{\Lambda\text{CDM}})$	$\log(z_{w_0w_a\text{CDM}})$	$\log \text{BF}$	Total computation time
CAMB + nested sampling	Unfeasible	Unfeasible	Unfeasible	12 years (projected, 48 CPUs)
CosmoPower-JAX + NUTS + harmonic	$406689.6^{+0.5}_{-0.3}$	$406687.7^{+0.5}_{-0.3}$	$1.9^{+0.7}_{-0.5}$	8 days (sampling, 24 GPUs) + 17 minutes (evidence, 1 GPU + 48 CPUs)
CosmoPower-JAX + NUTS + naïve flow estimator	406703 ± 39	406701 ± 62	2 ± 73	Similar to harmonic

Defining required accuracy

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

$$\epsilon = \sqrt{\frac{1}{N_\nu} \sum_i^{N_t} (S_{\text{true}}(t) - S_{\text{pred}}(t))^2}$$

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



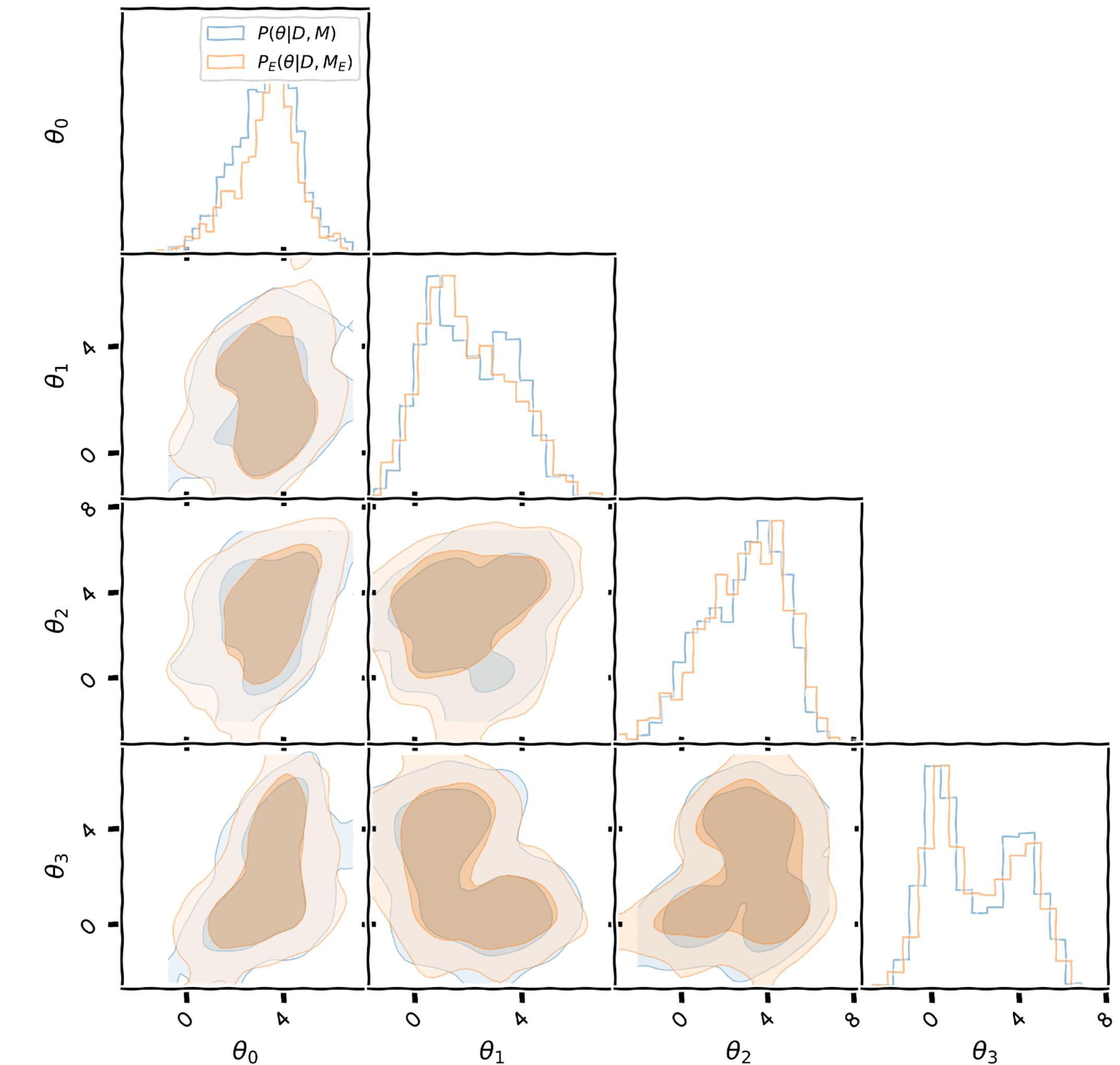
Impact on posterior recovery?

- Really interested in is how well can we recover the posteriors if we use an emulator rather than the full simulation?

$$\log L \rightarrow \log L + \delta \log L$$

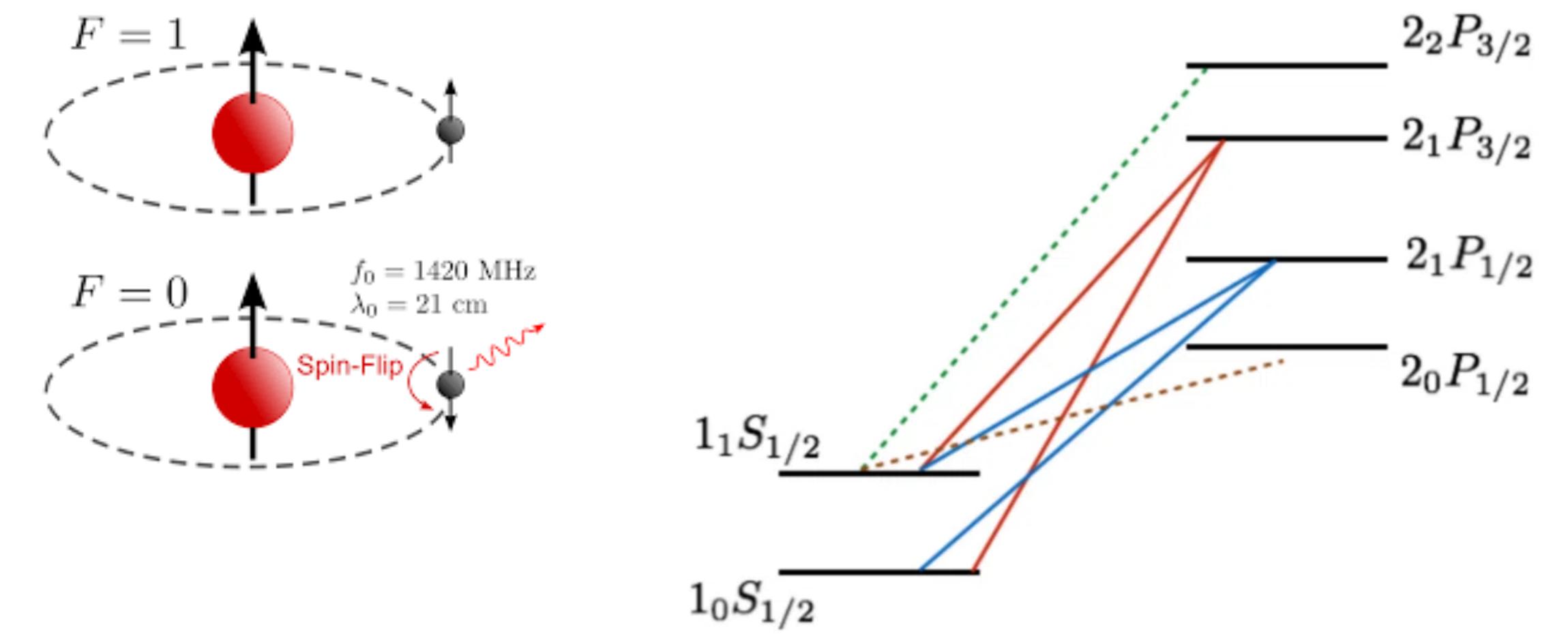
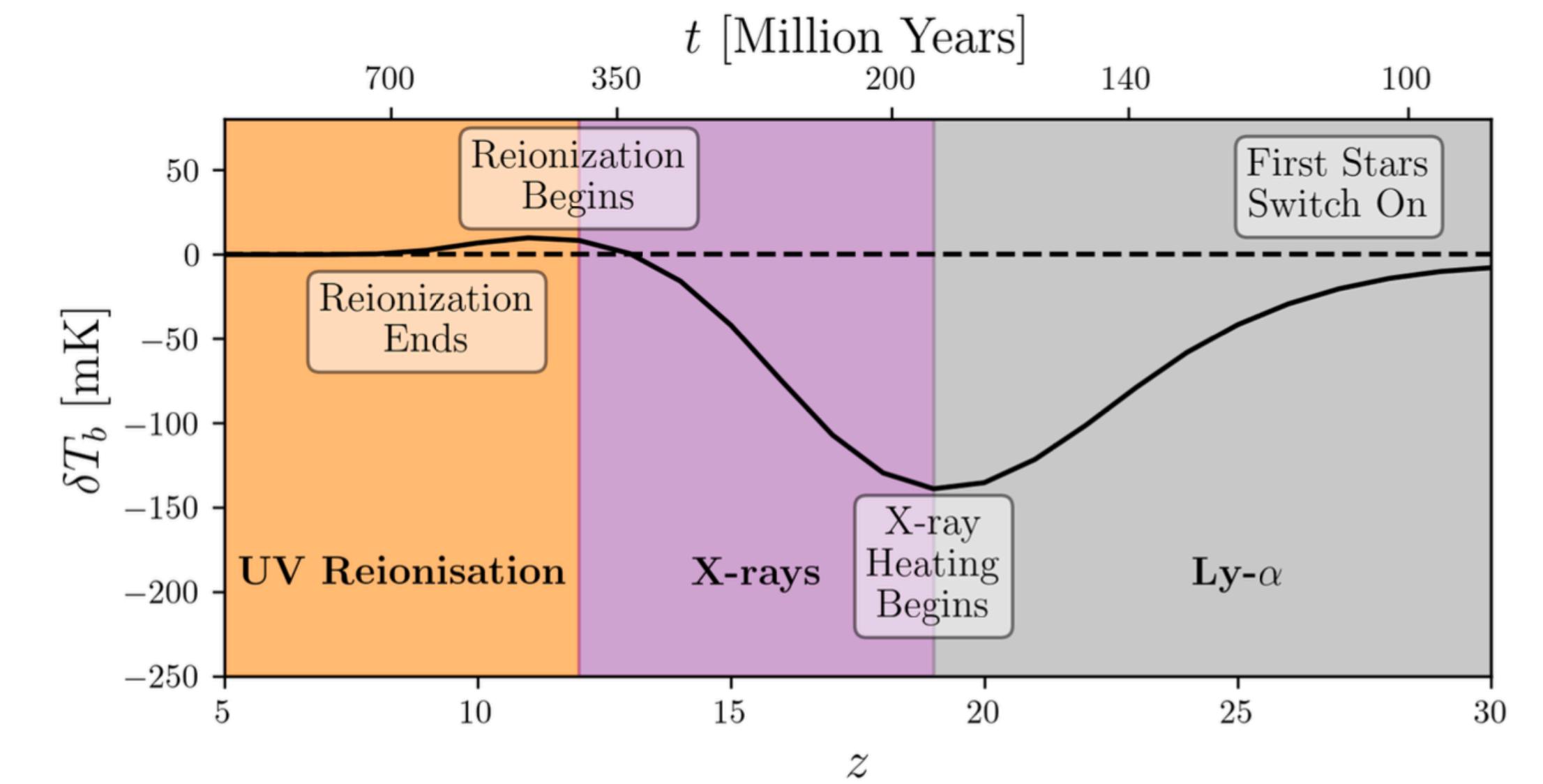
$$P(\theta|D, M) = \frac{L\pi}{\int L\pi d\theta} \rightarrow P_E(\theta|D, M_E) = \frac{L\pi e^{\delta \log L}}{\int L\pi e^{\delta \log L} d\theta}$$

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



21cm Cosmology

- Relative brightness of 21cm signal from neutral hydrogen and the background CMB
- 21cm signal brightness measured by a statistical temperature
- Relative number of atoms with aligned and anti-aligned proton and electron spins driven by many different processes
 - Cosmology ($z < 30$)
 - Star formation ($30 < z < 15$)
 - X-ray heating ($15 < z < 8$)
 - Ionisation ($8 < z < 5$)
 - With some overlap
 - And many other processes



Dorigo Jones+23

- Inference with a 1D radiative transfer code ARES and compared recovered posteriors to posteriors recovered with an emulator of ARES
- Want to use semi-numerical or hydro simulations
- Measured posterior accuracy with two metrics

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

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

- They concluded that can't accurately recover the posteriors with an emulator with $\bar{\epsilon} \approx 0.05\sigma$

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

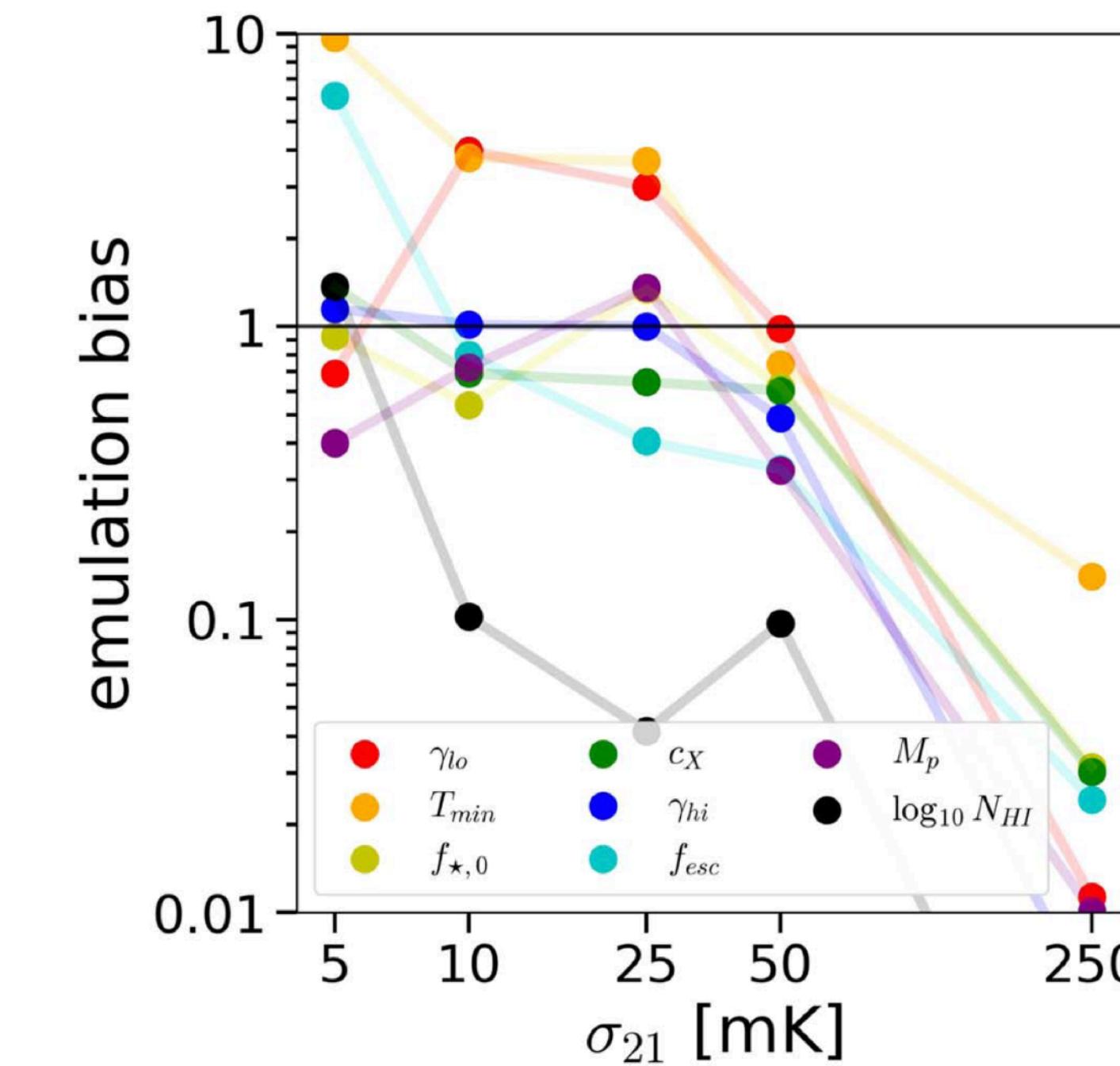
J. Dorigo Jones¹ , D. Rapetti^{1,2,3} , J. Mirocha^{4,5} , J. J. Hibbard¹ , J. O. Burns¹ , and N. Bassett¹ 

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

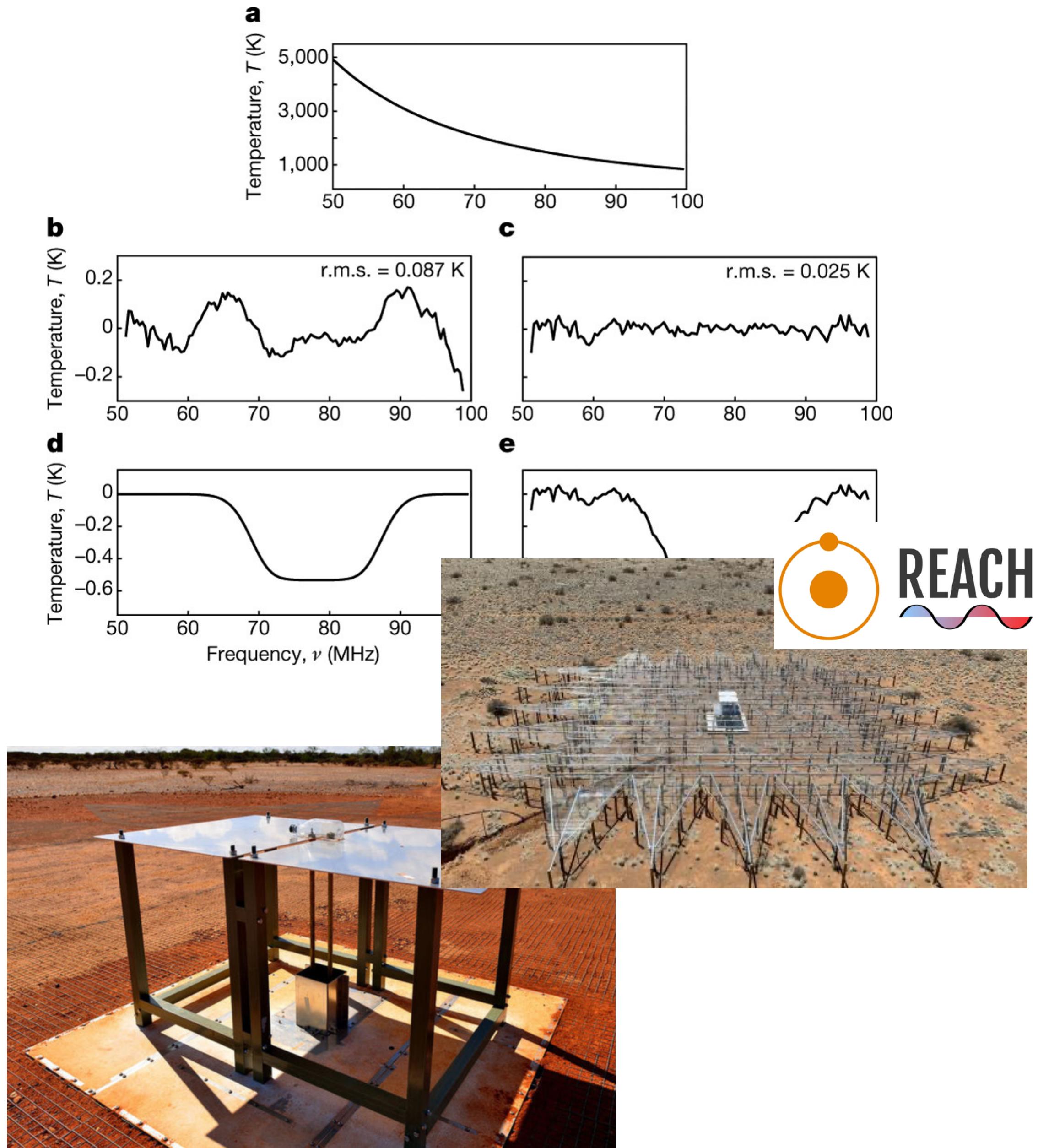


Why this is concerning?

- We need to go down to around 25 mK noise to confidently detect the 21cm signal
- Most emulators have $\bar{\epsilon} \approx 1 \text{ mK} \approx 0.05 \times 25\text{mK}$ and it seems challenging to go beyond this
- If we assume a Gaussian likelihood and

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

we would expect the uncertainty from the instrument to dominate the posteriors

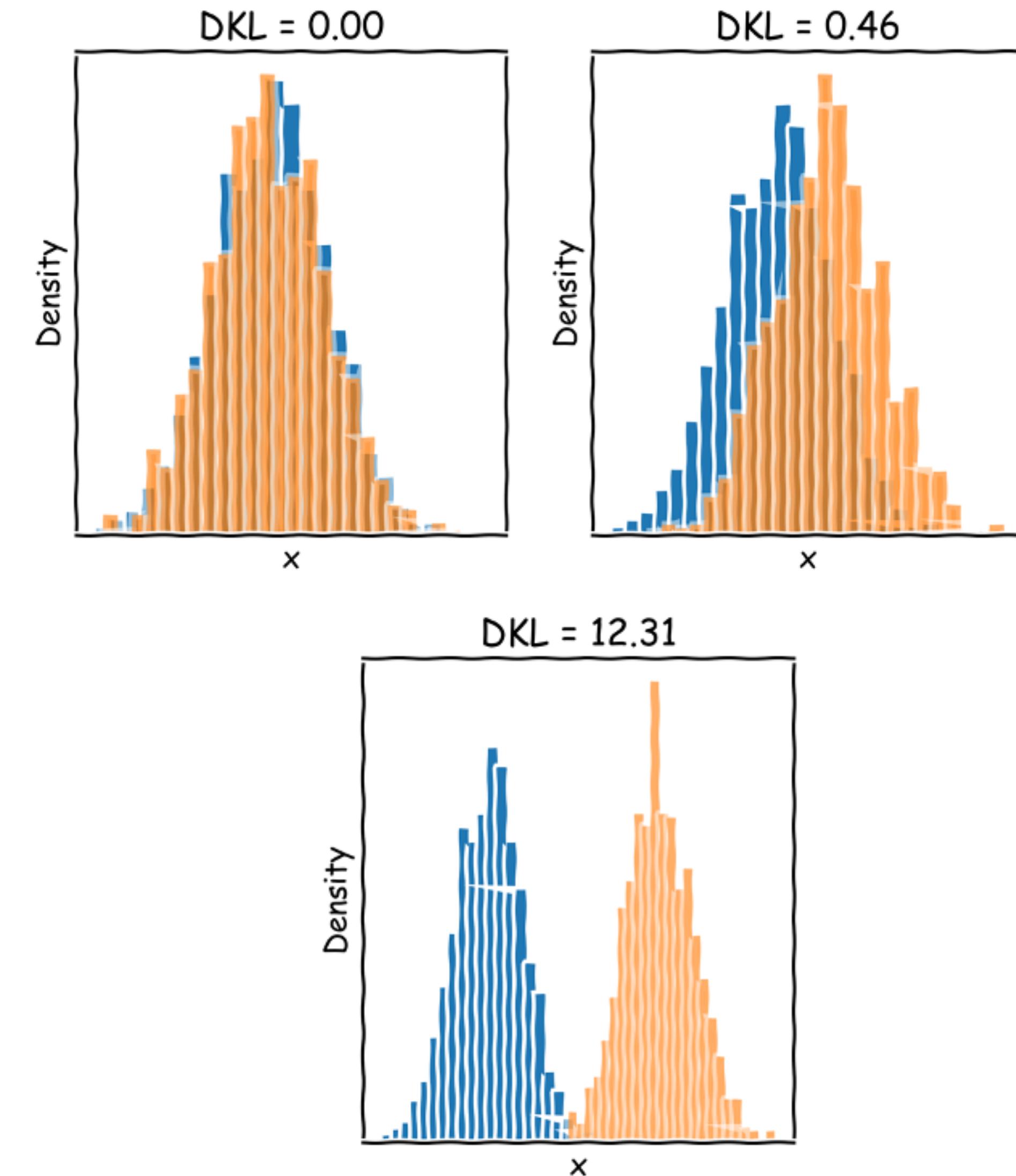


Measuring the impact of the emulator

- The emulator bias defined in Dorigo Jones+23 is fine but its only really considers the difference in 1D
- More comprehensive measure of the difference between the true and emulated posteriors is the Kullback-Leibler Divergence

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

- Typically do not have access to P else we wouldn't be interested in emulators



Measuring the impact of the emulator

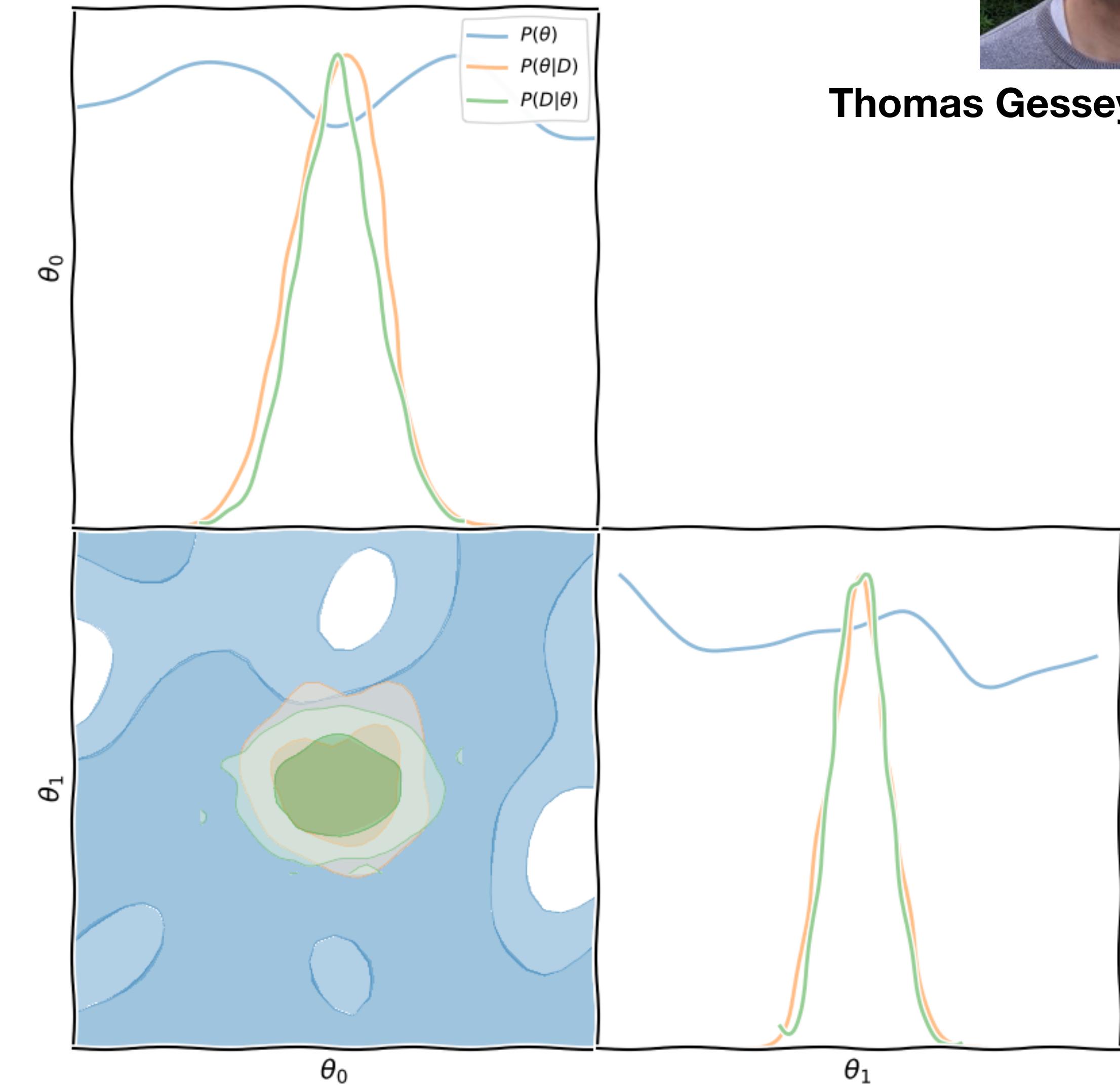


Thomas Gessey-Jones

- Can make progress if we make some assumptions
- Firstly we assume that the likelihood function is Gaussian

$$L \propto \exp\left(-\frac{1}{2}(D - \mathcal{M})^T \Sigma^{-1} (D - \mathcal{M})\right)$$

And our prior is uniform such that the posterior is also Gaussian



Measuring the impact of the emulator

- Assume a linear model and linear emulator error

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

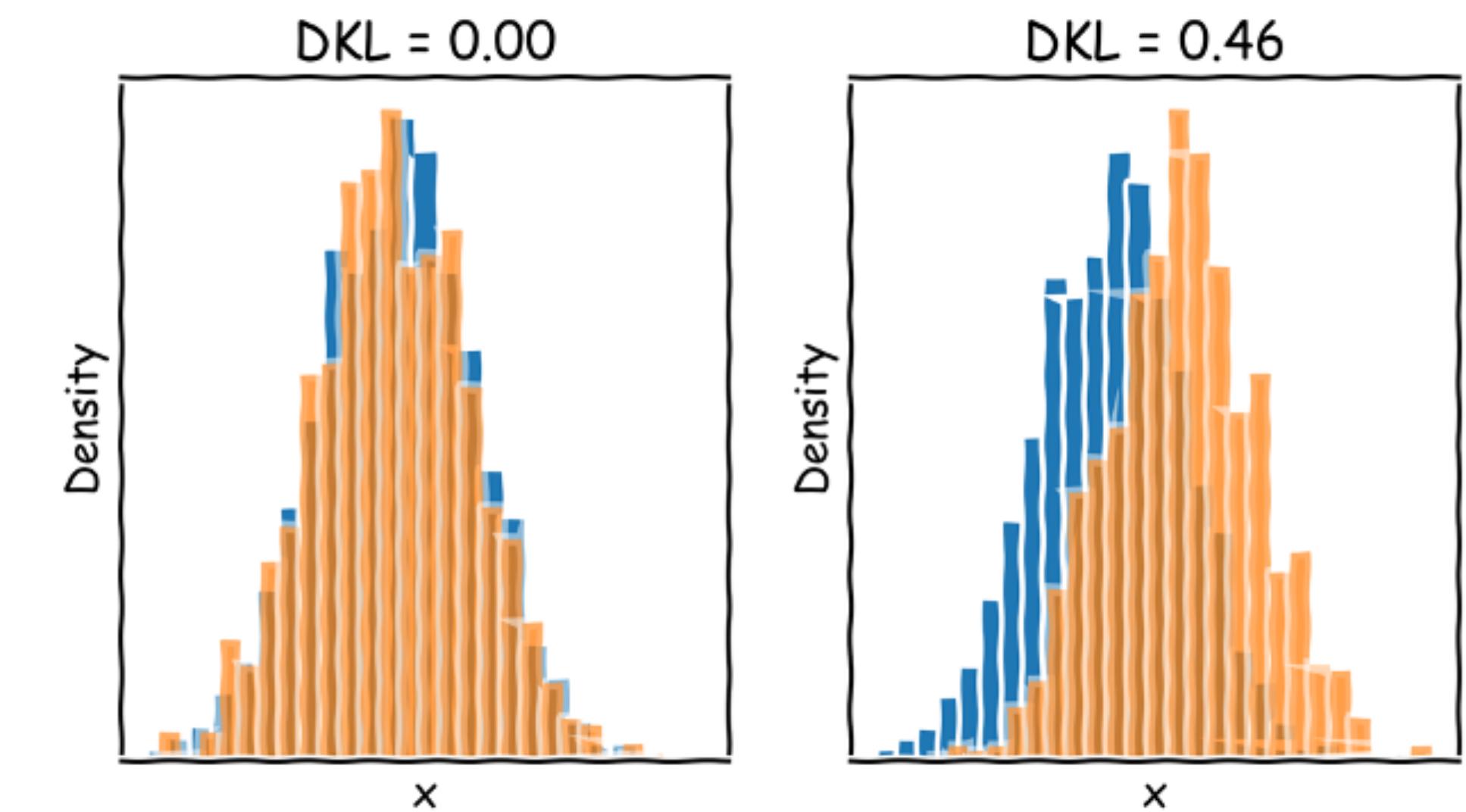
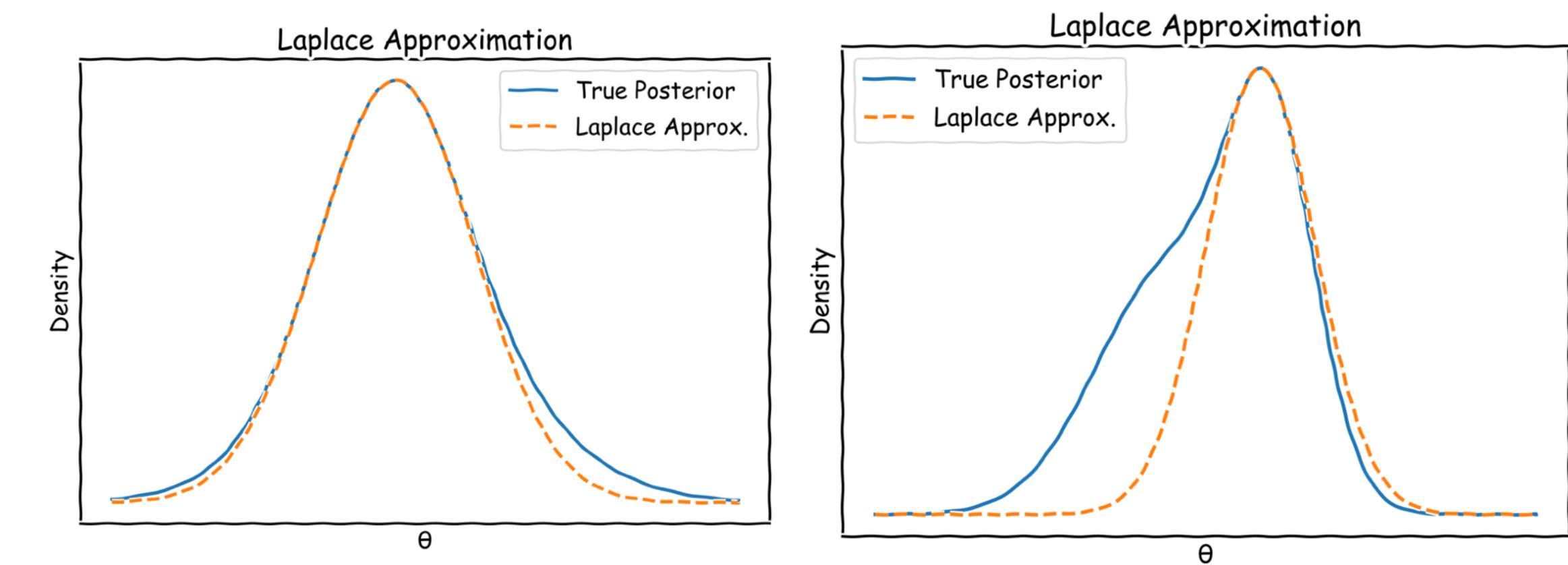
Such that $M_e(\theta) = (M + E)\theta + (m + \epsilon)$

- Comes from Taylor expansion of model around the MAP and the assumption that the posterior is sharply peaked

- KL divergence between two Gaussians P and P_E is given by

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

- Make assumptions about $E \ll M$ and $\Sigma = \frac{1}{\sigma^2} \mathbf{1}_{N_d}$

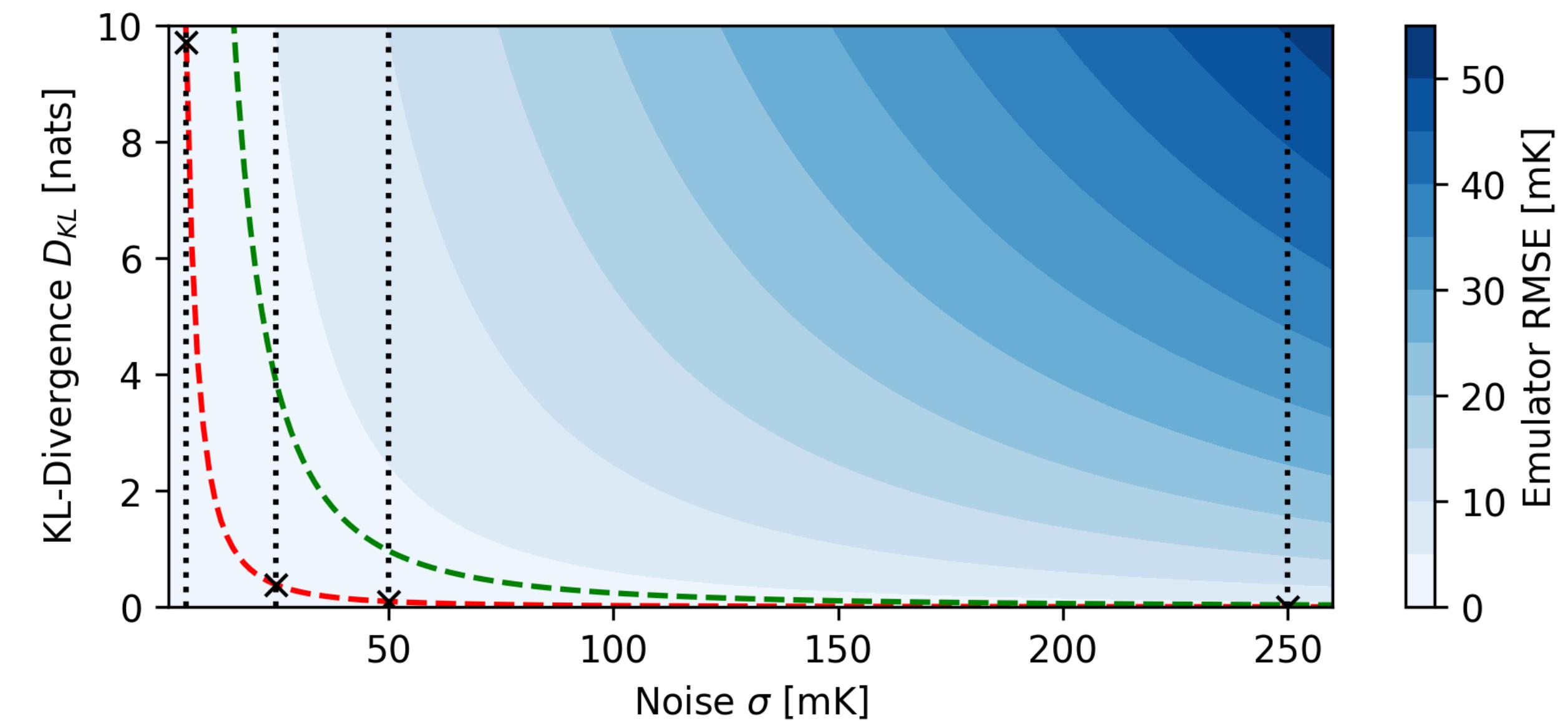


Measuring the impact of the emulator

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

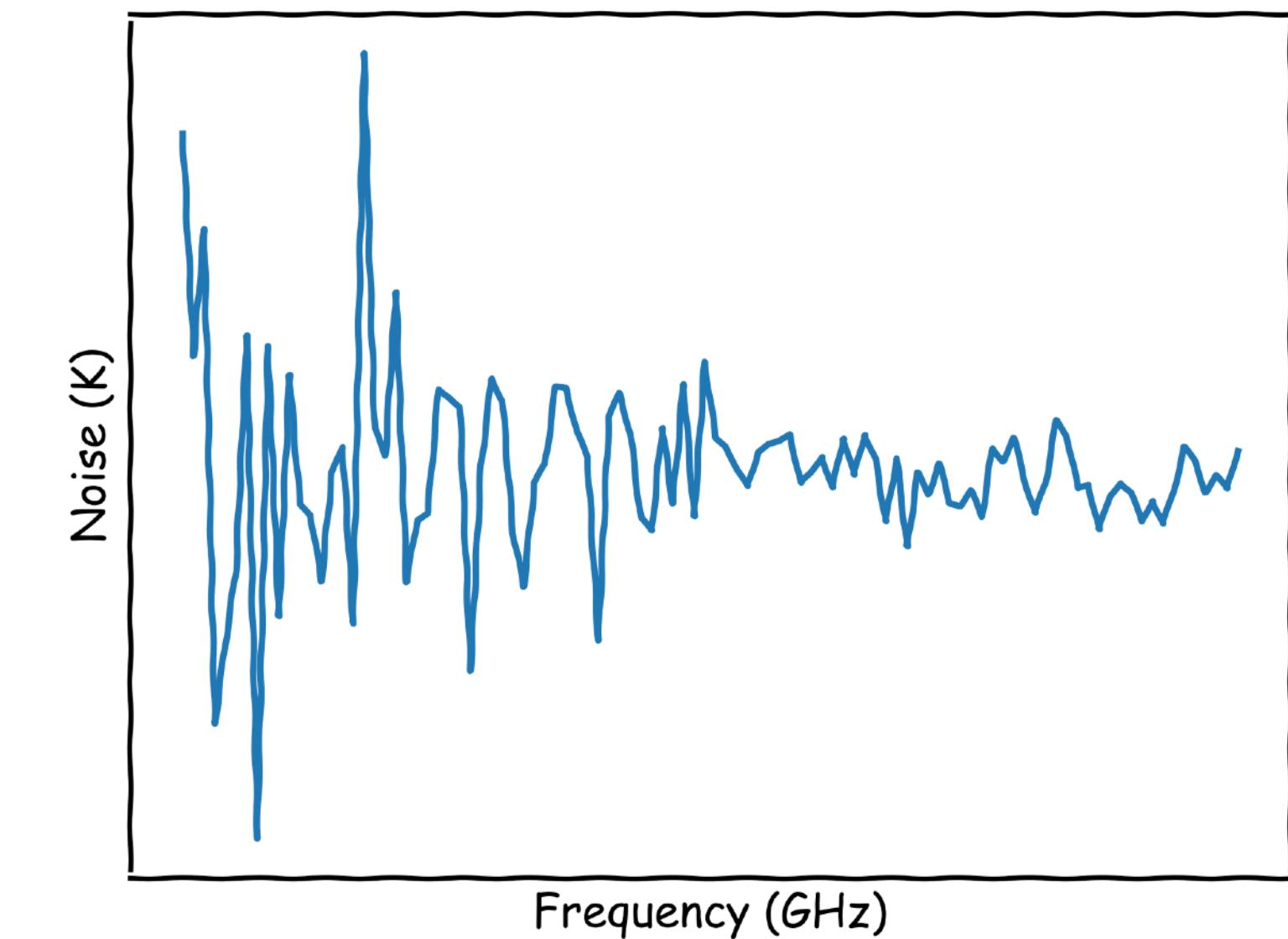
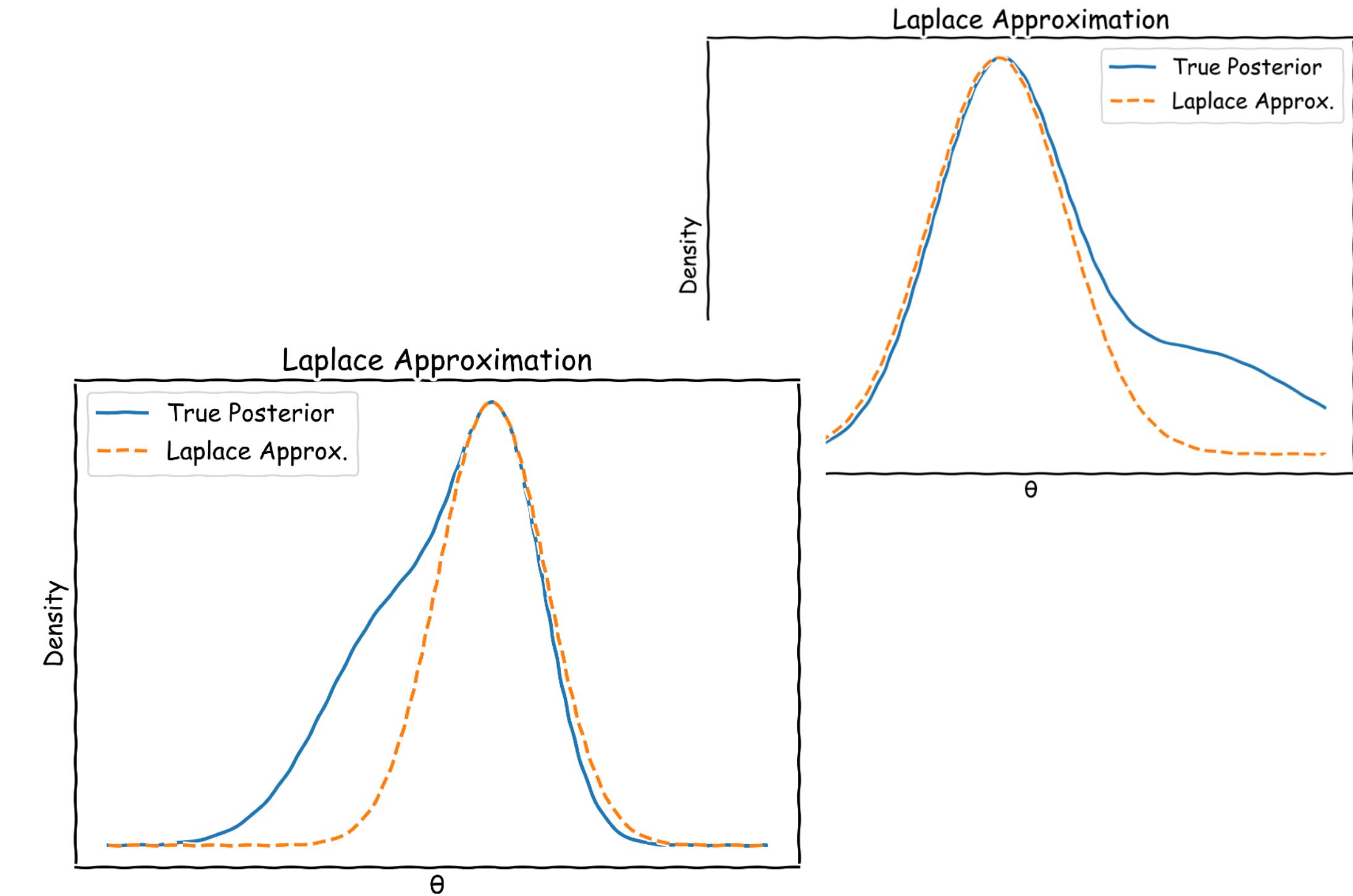
$$D_{\text{KL}}(P || P_\epsilon) \leq \frac{N_d}{2} \left(\frac{\text{RMSE}}{\sigma} \right)^2$$

- Function of 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



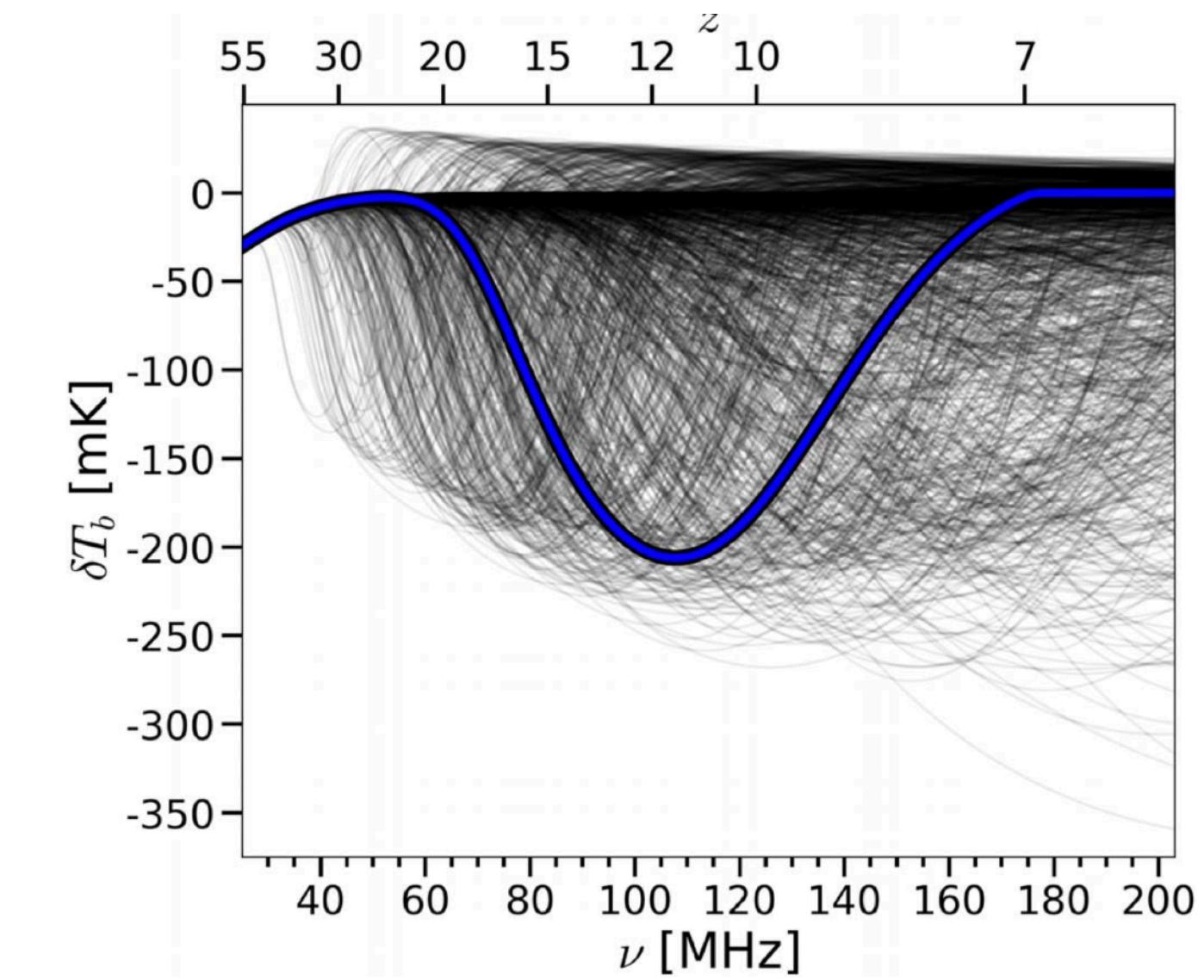
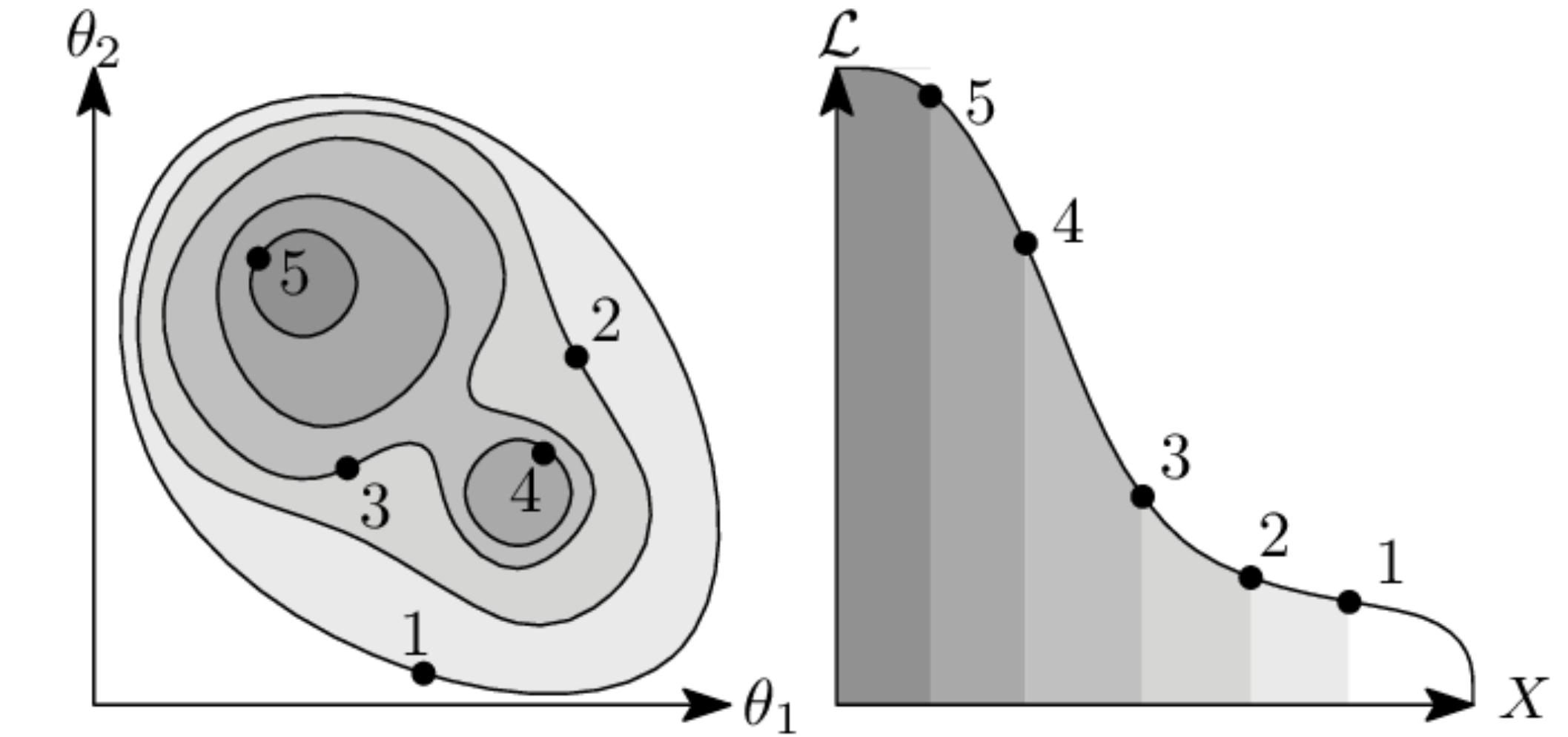
Limitations of the approximation

- 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

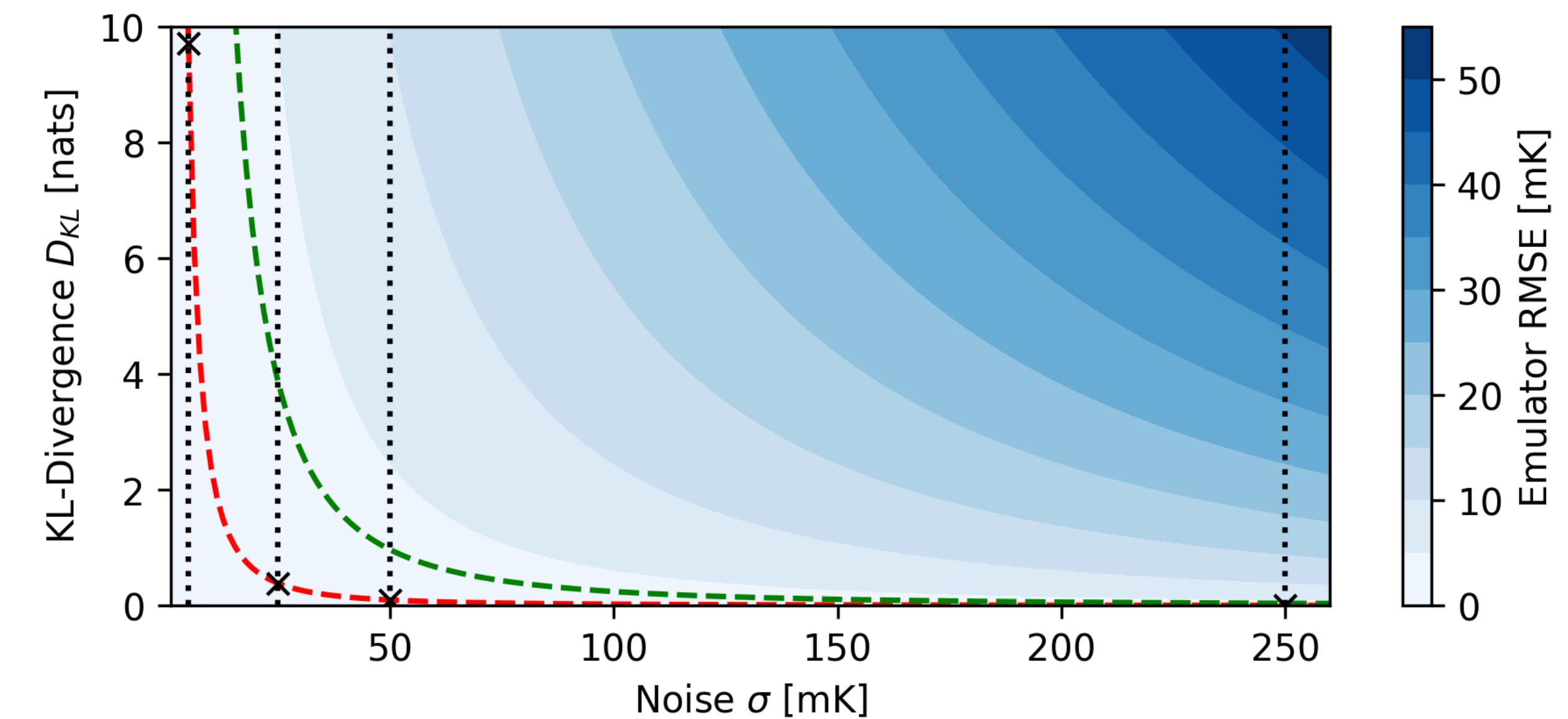
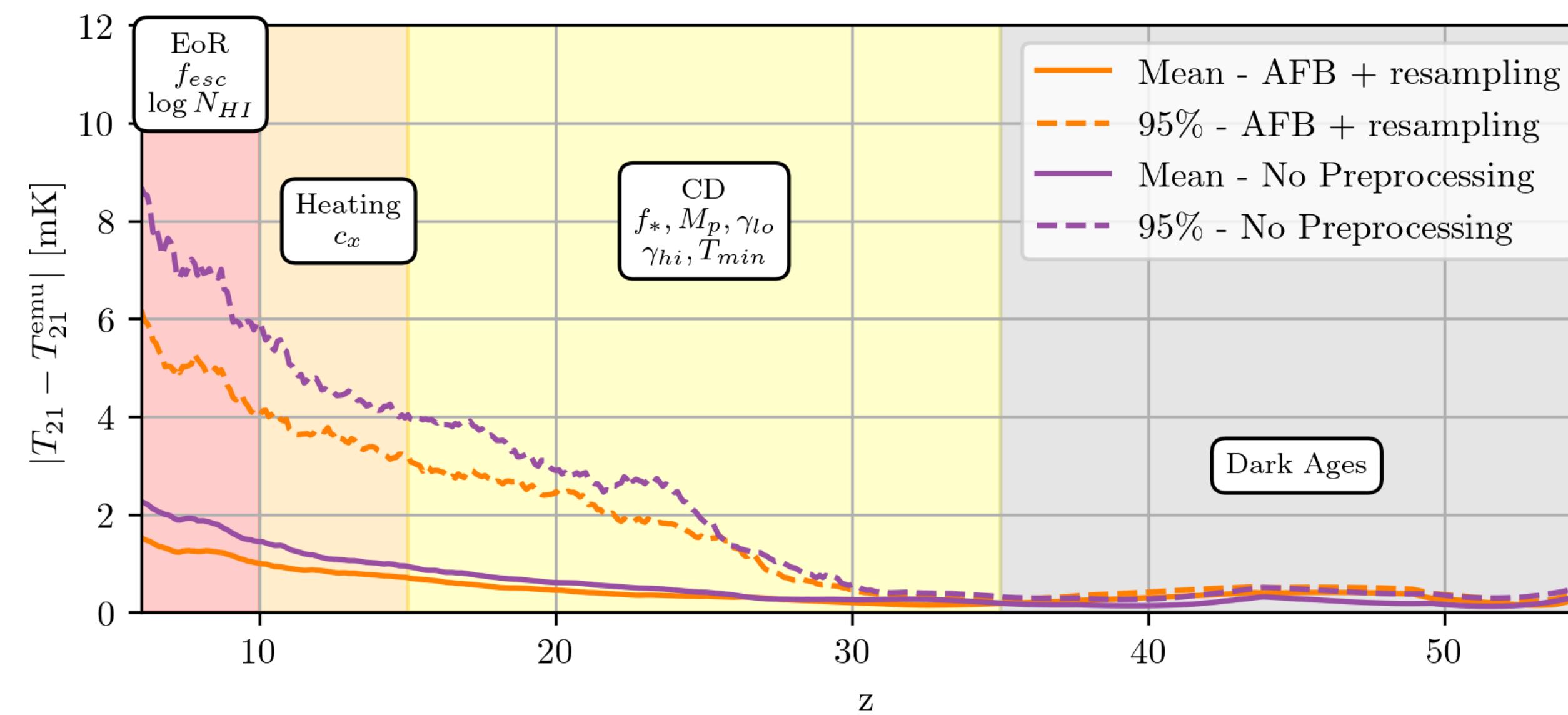


Testing on a 21cm Cosmology problem

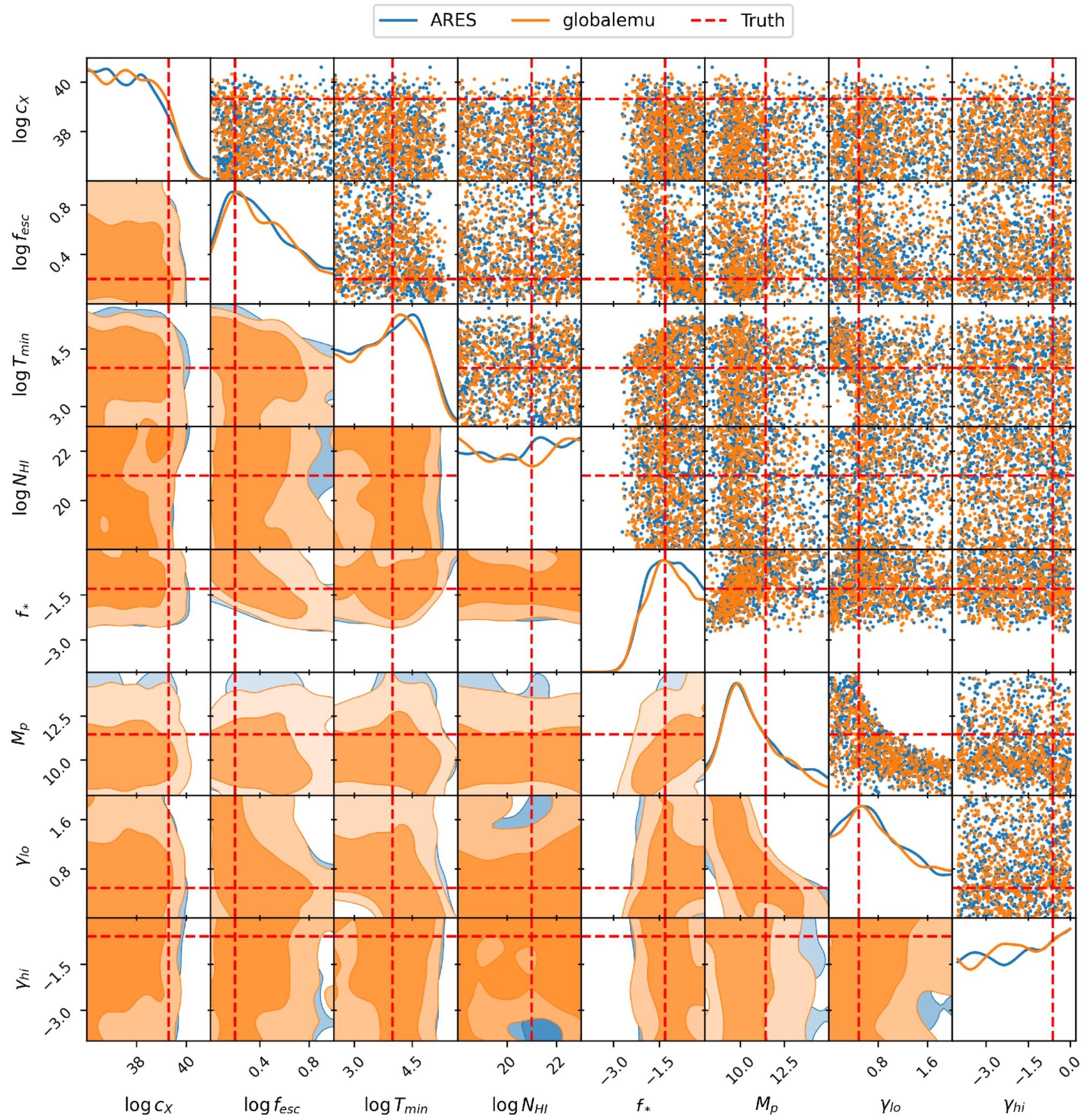
- Assuming the data comprises of signal plus noise
- Same fiducial signal as in Dorigo Jones+23
- Same prior range and same sampler
- Assuming a Gaussian likelihood as was done in their paper
- Assuming absolute knowledge of the level of noise in the data
- Running for 5, 25, 50 and 250 mK



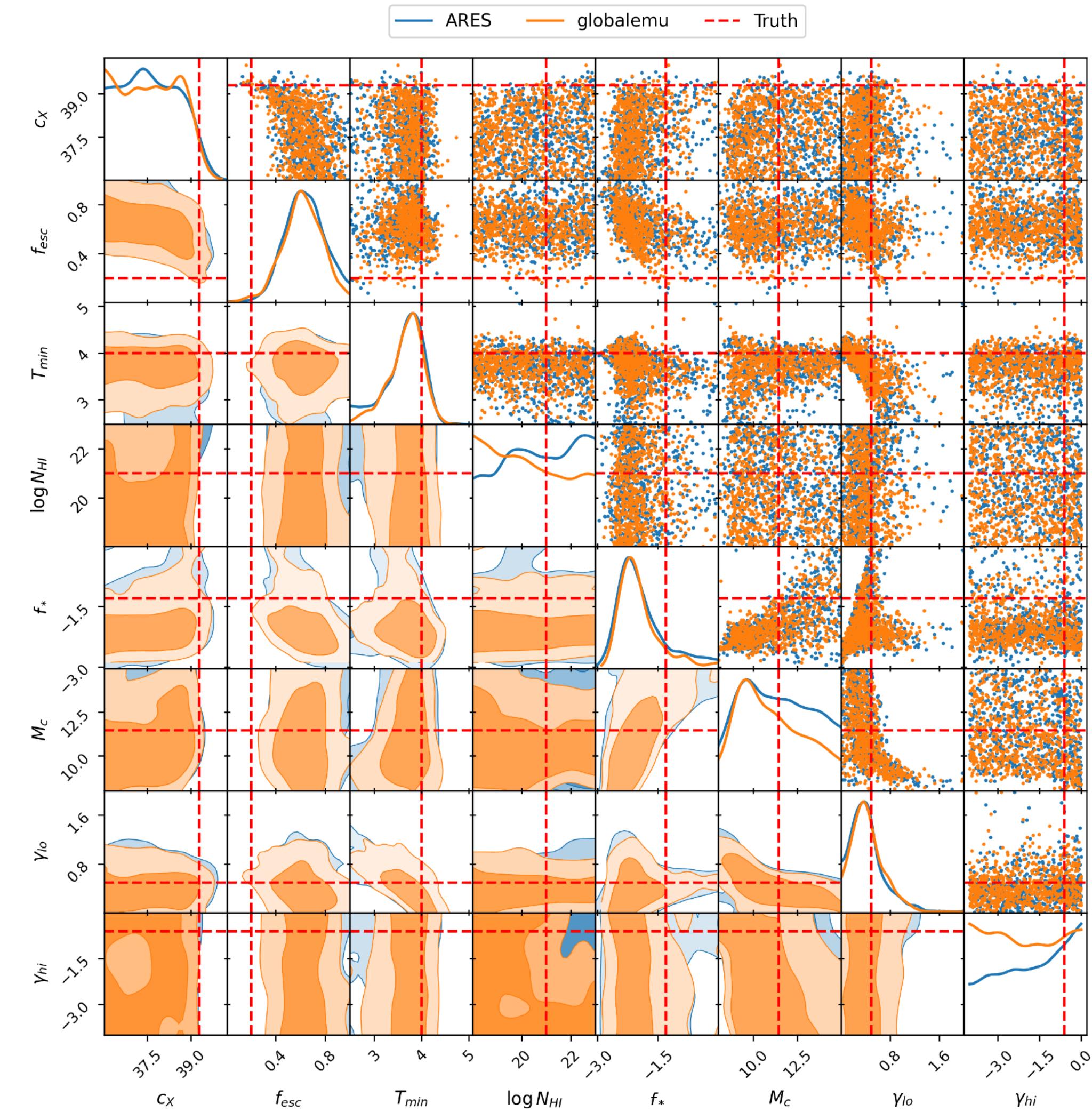
globalemu performance and ARES modelling



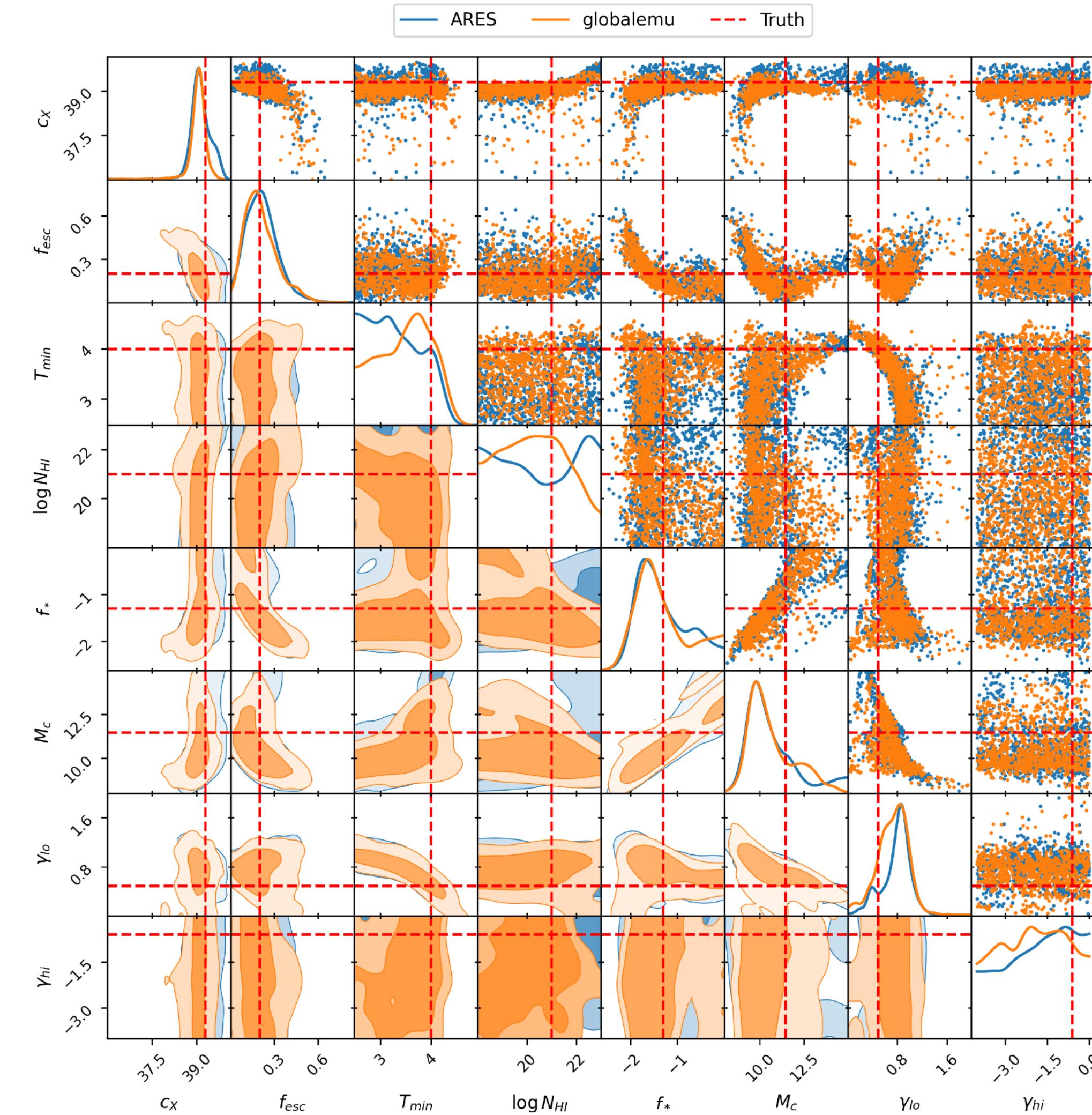
Running the analysis - 250 mK



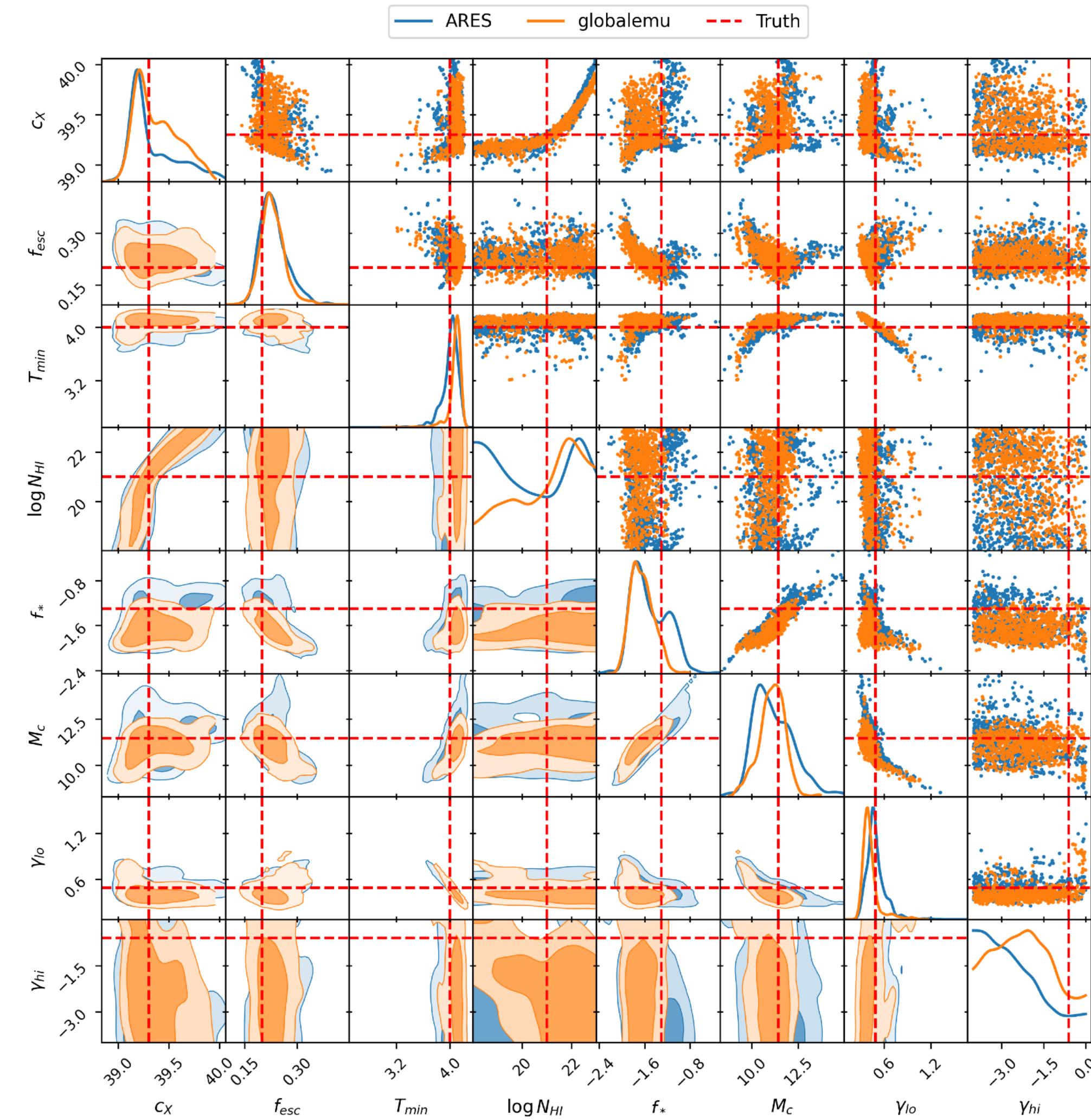
Running the analysis - 50 mK



Running the analysis - 25 mK

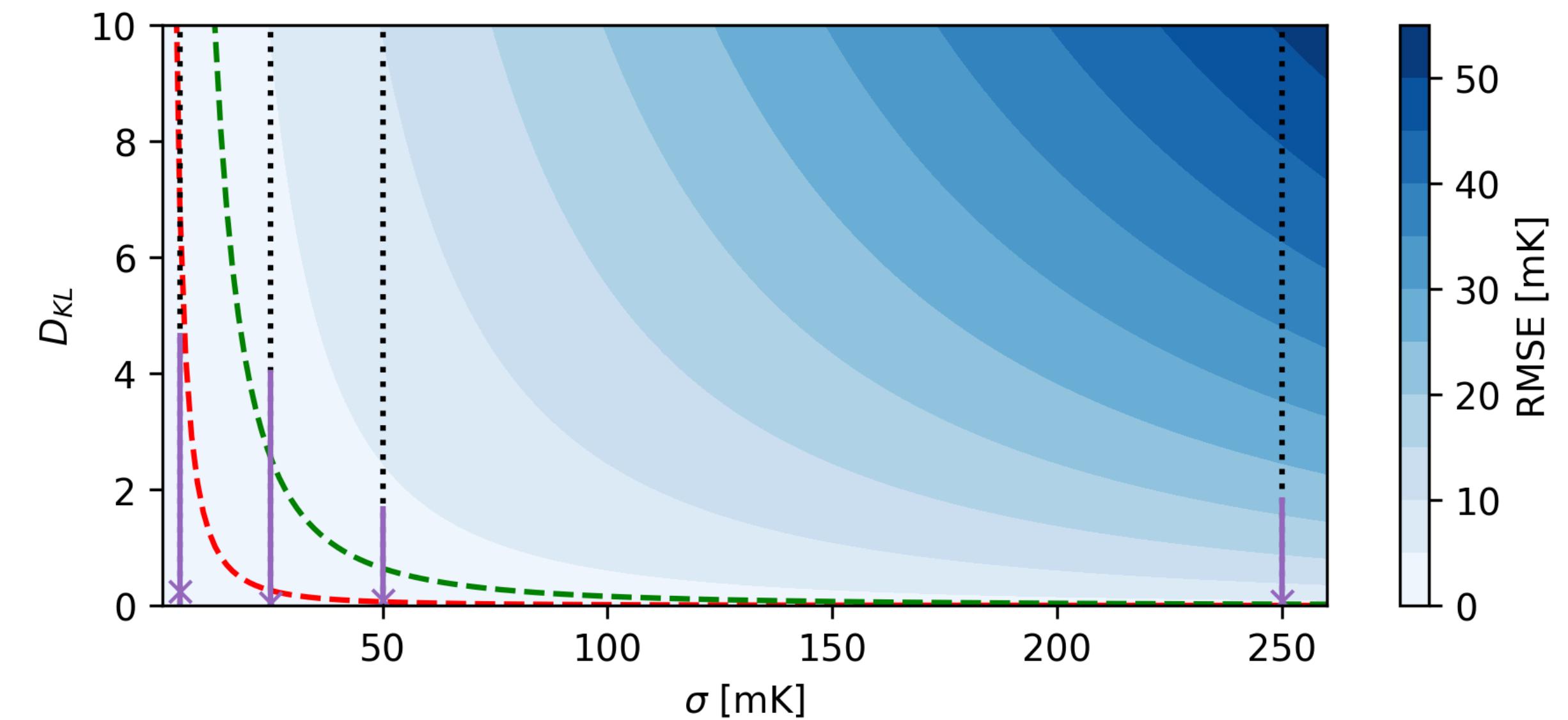


Running the analysis - 5 mK



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* [see Bevins et al 2022, 2023, arXiv:2207.11457, arXiv:2205.12841]
- Compare calculated D_{KL} with predicted upper limits



Noise Level [mK]	Estimated $\mathcal{D}_{KL} \leq$		Actual \mathcal{D}_{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	$0.08^{+1.78}_{-0.02}$

Conclusions

- 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 $\bar{\epsilon} \approx 0.2\sigma$ for 21cm
- arXiv:2503.13263
- https://github.com/htjb/validating_posteriors

