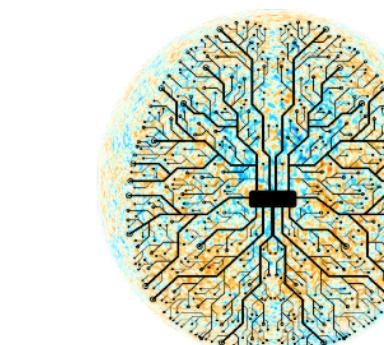
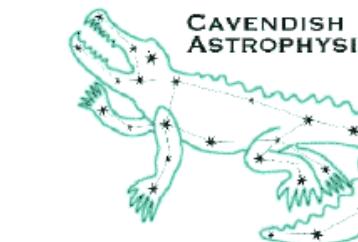


On the accuracy of posterior recovery with neural networks

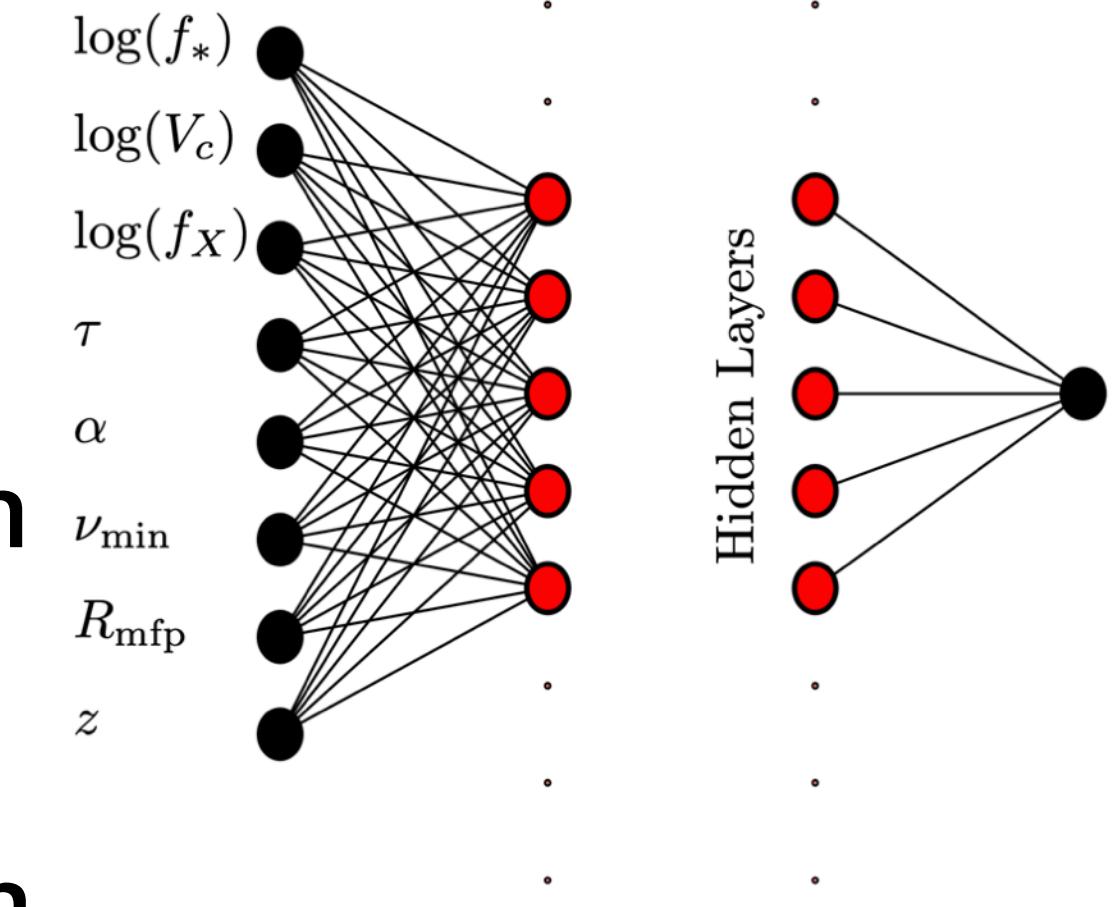
Harry Bevins

Thomas Gessey-Jones, Will Handley

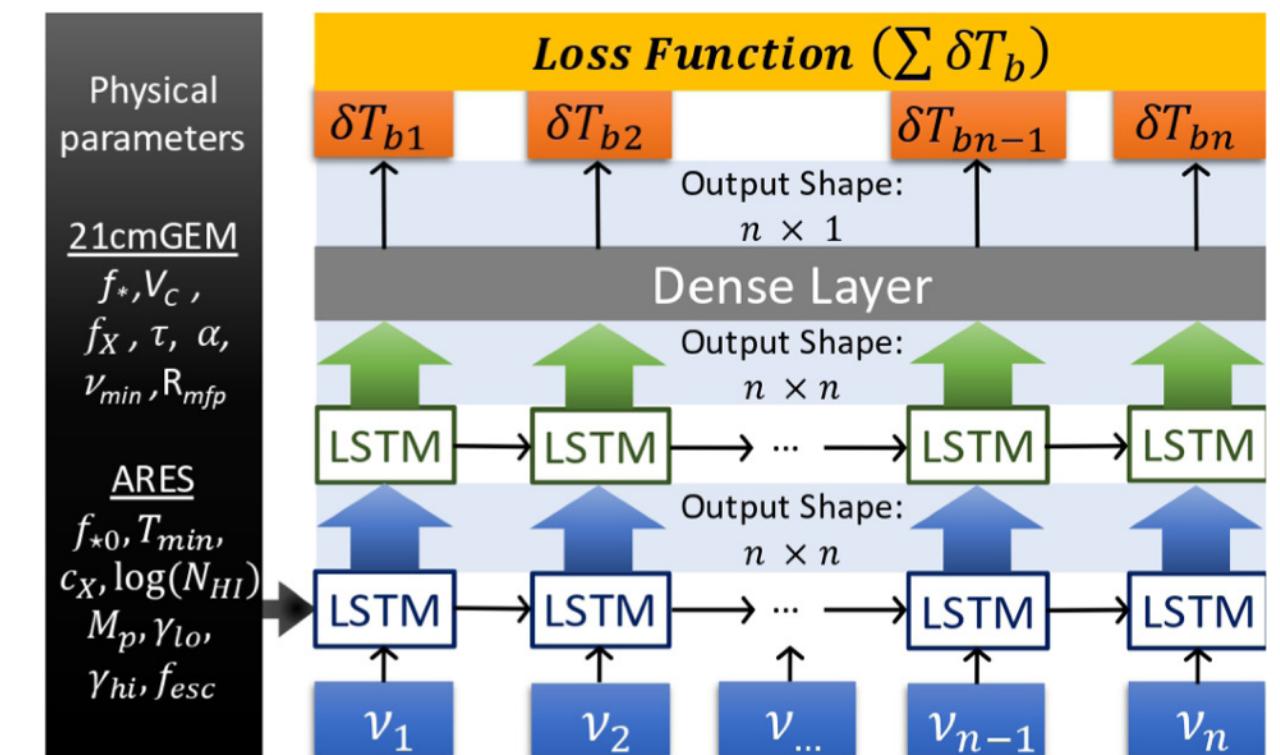
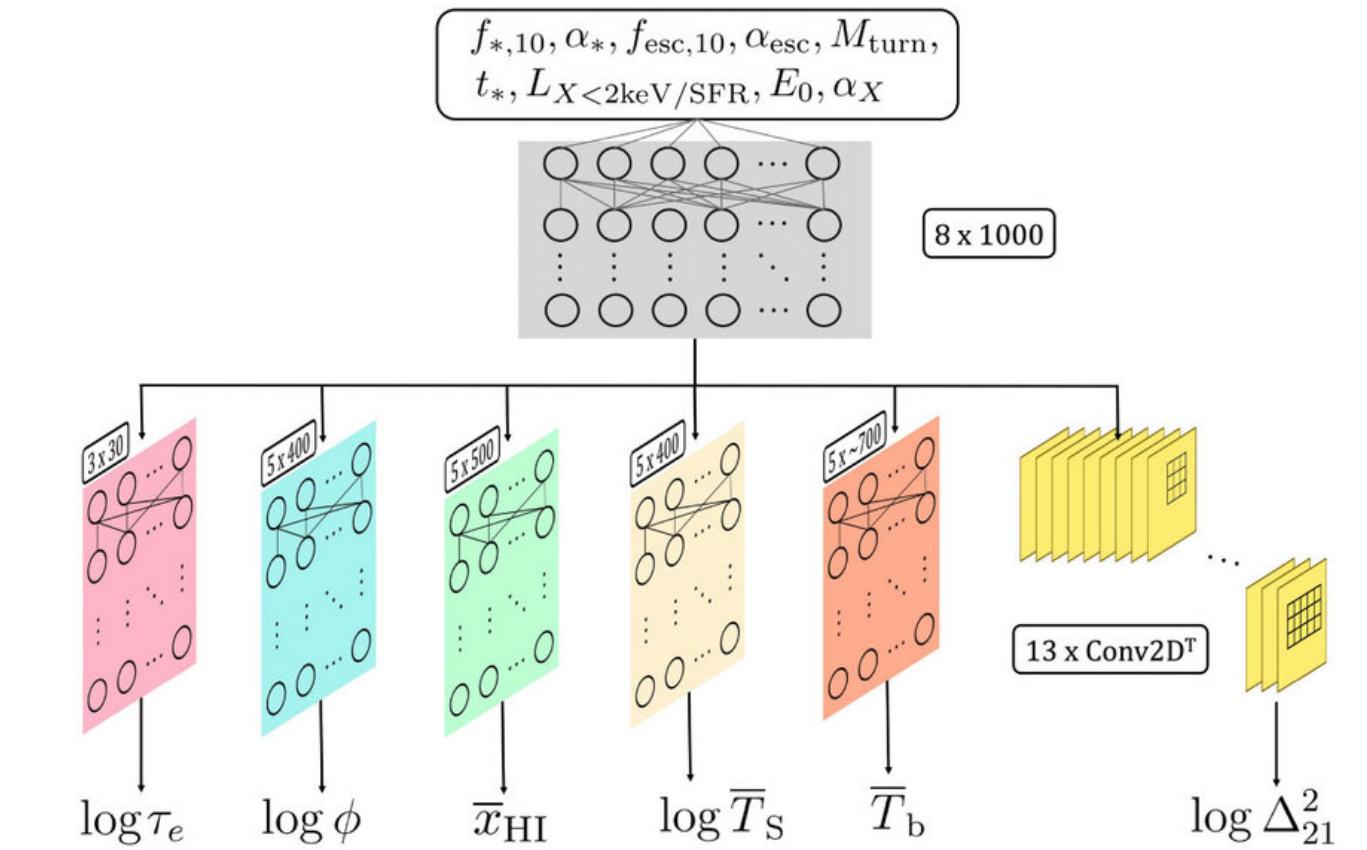
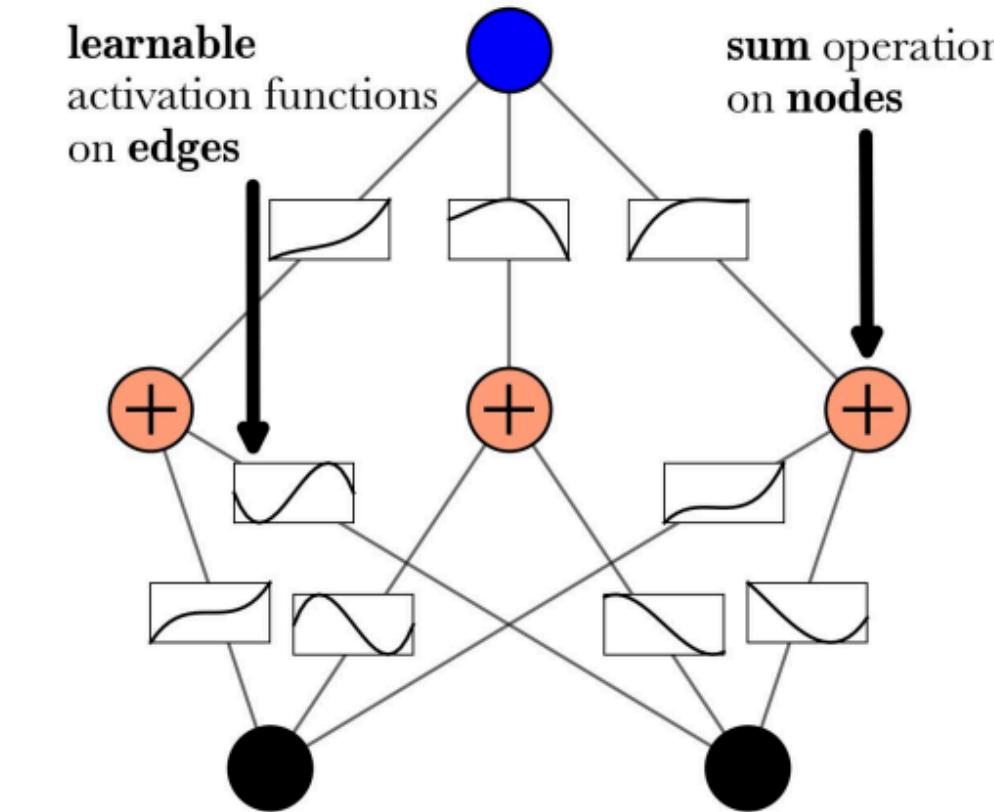


Emulators in 21cm Cosmology

- 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]



Kolmogorov-Arnold Network (KAN)

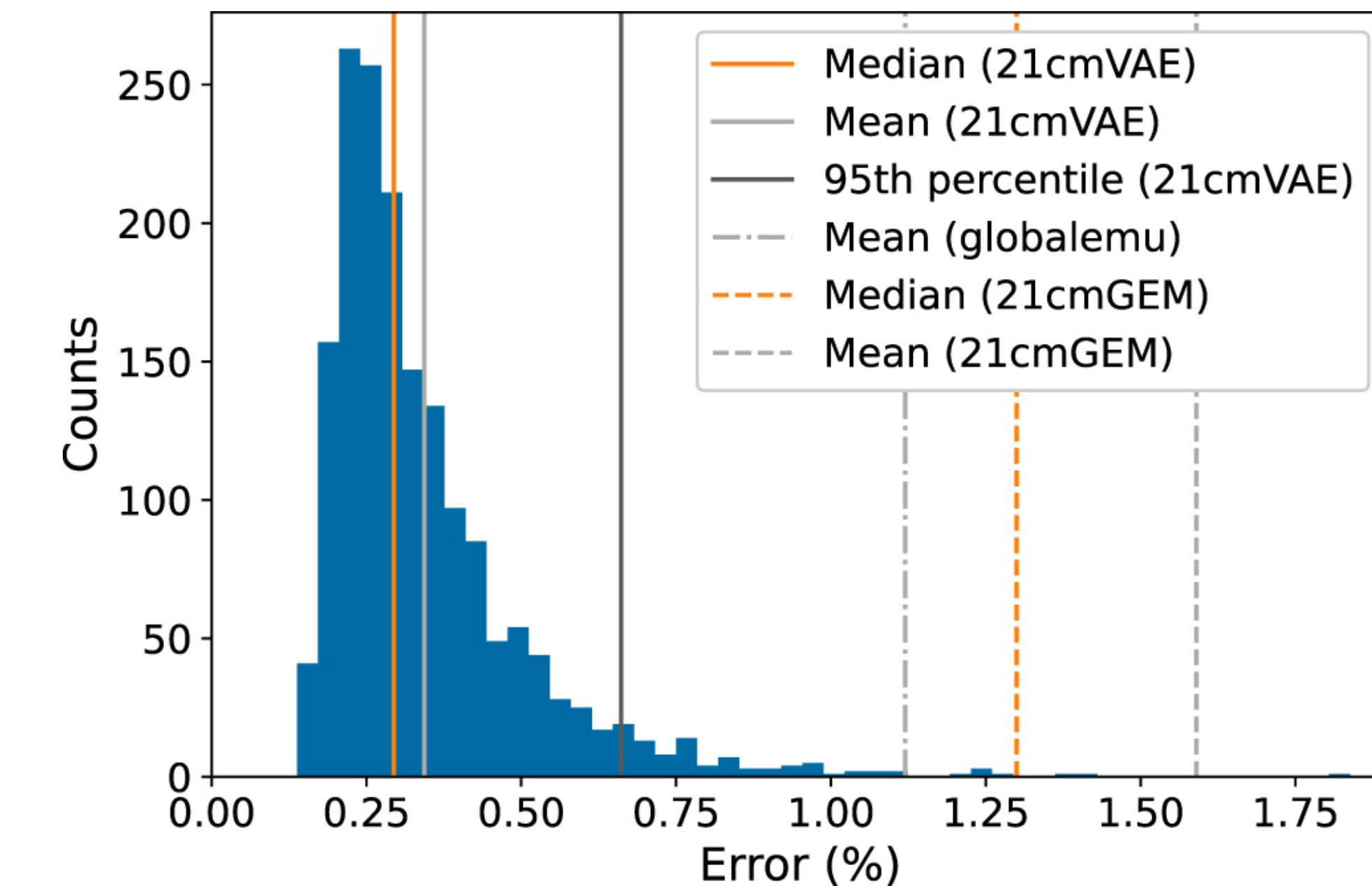


Defining required accuracy

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

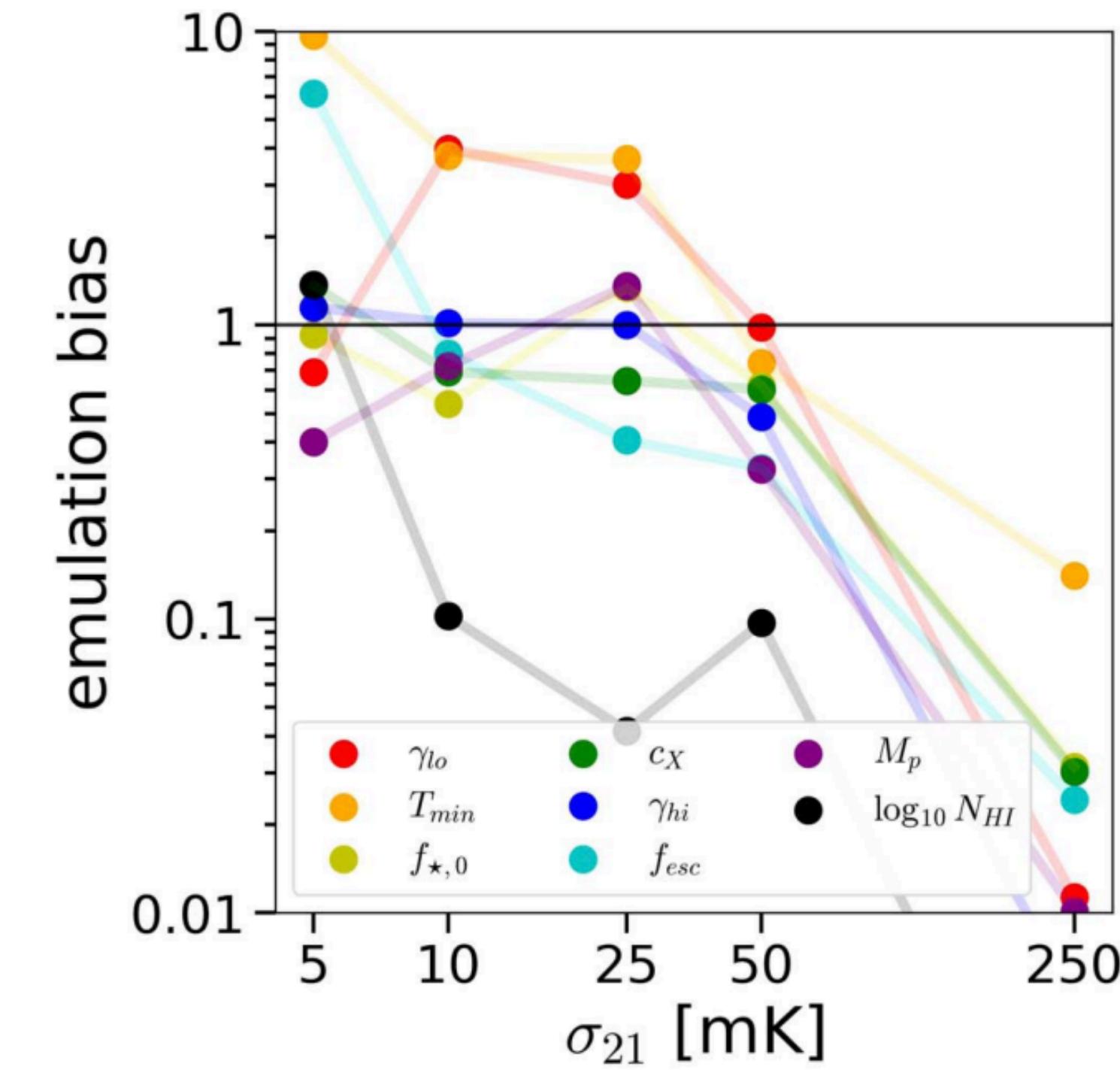
$$\epsilon = \sqrt{\frac{1}{N_\nu} \sum_i^{N_t} (T_{\text{true}}^{21}(z) - T_{\text{pred}}^{21}(t))^2}$$

- 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 $\bar{\epsilon} \lesssim 0.1\sigma$



Defining required accuracy

- Dorigo Jones et al. 2023 started to ask and answer this question
- Making a direct comparison between $P(\theta | D, M) \leftrightarrow P(\theta | D, M_E)$
- We wanted to come up with something more predictive because we don't have $P(\theta | D, M)$ only $P(\theta | D, M_E)$
- “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?

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

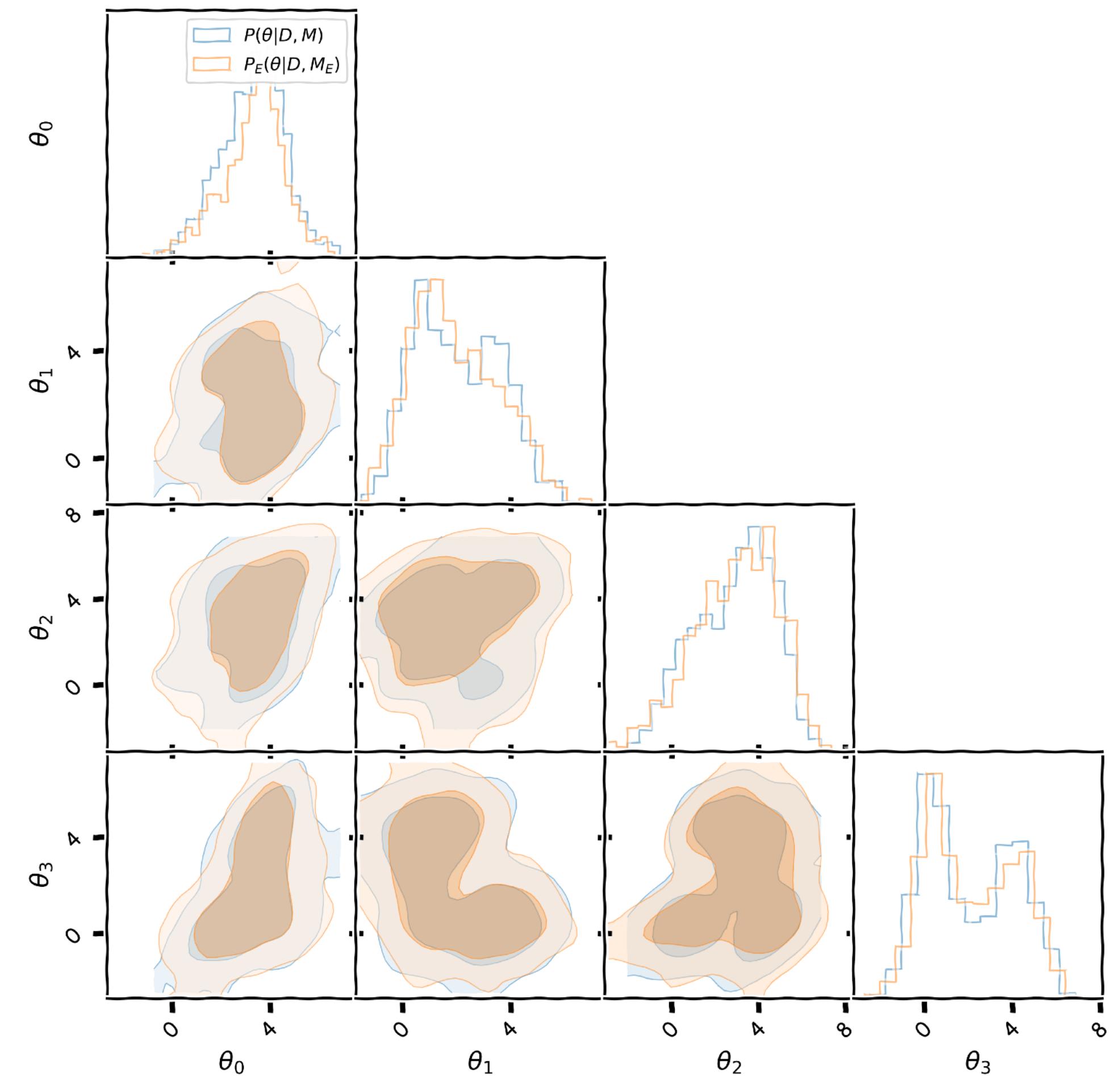
$$P = \frac{L}{Z} \pi$$

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

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

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

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

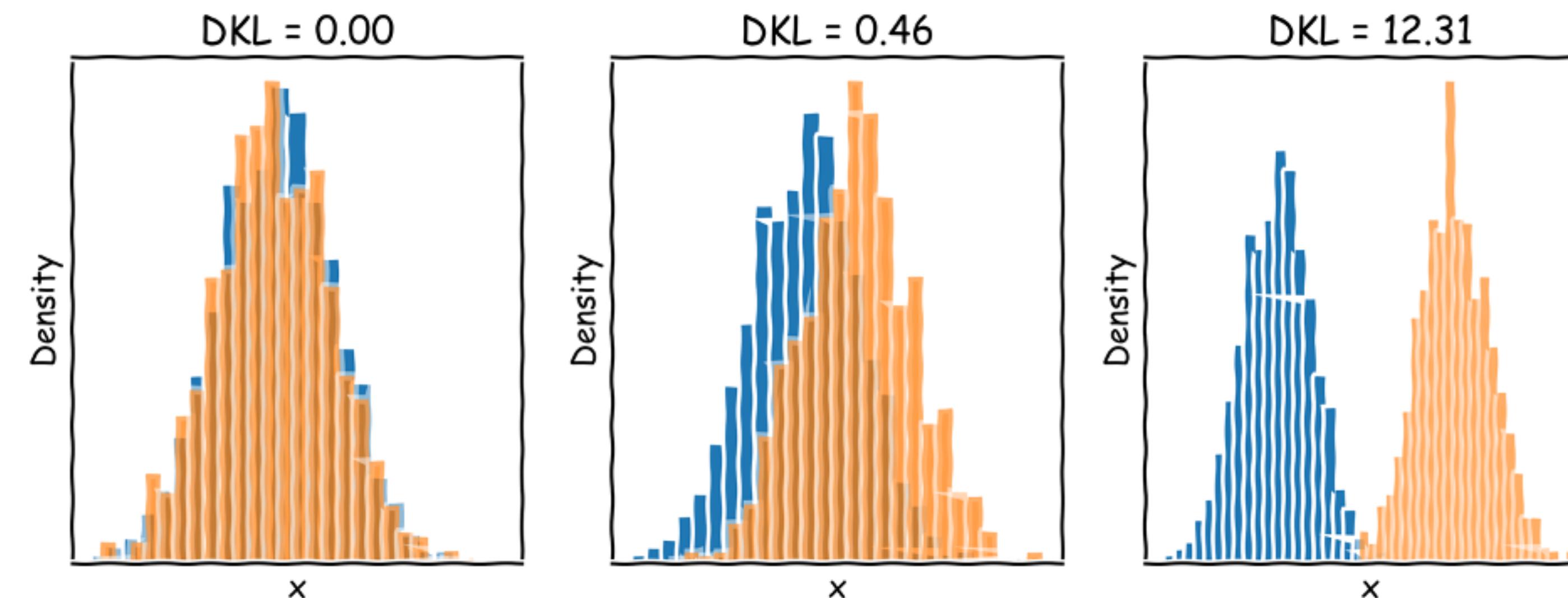


Measuring the impact of the emulator

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

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

- 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 $D_{KL}(P || P_E)$

$$L = \mathcal{N}(D; \Sigma, M(\theta))$$

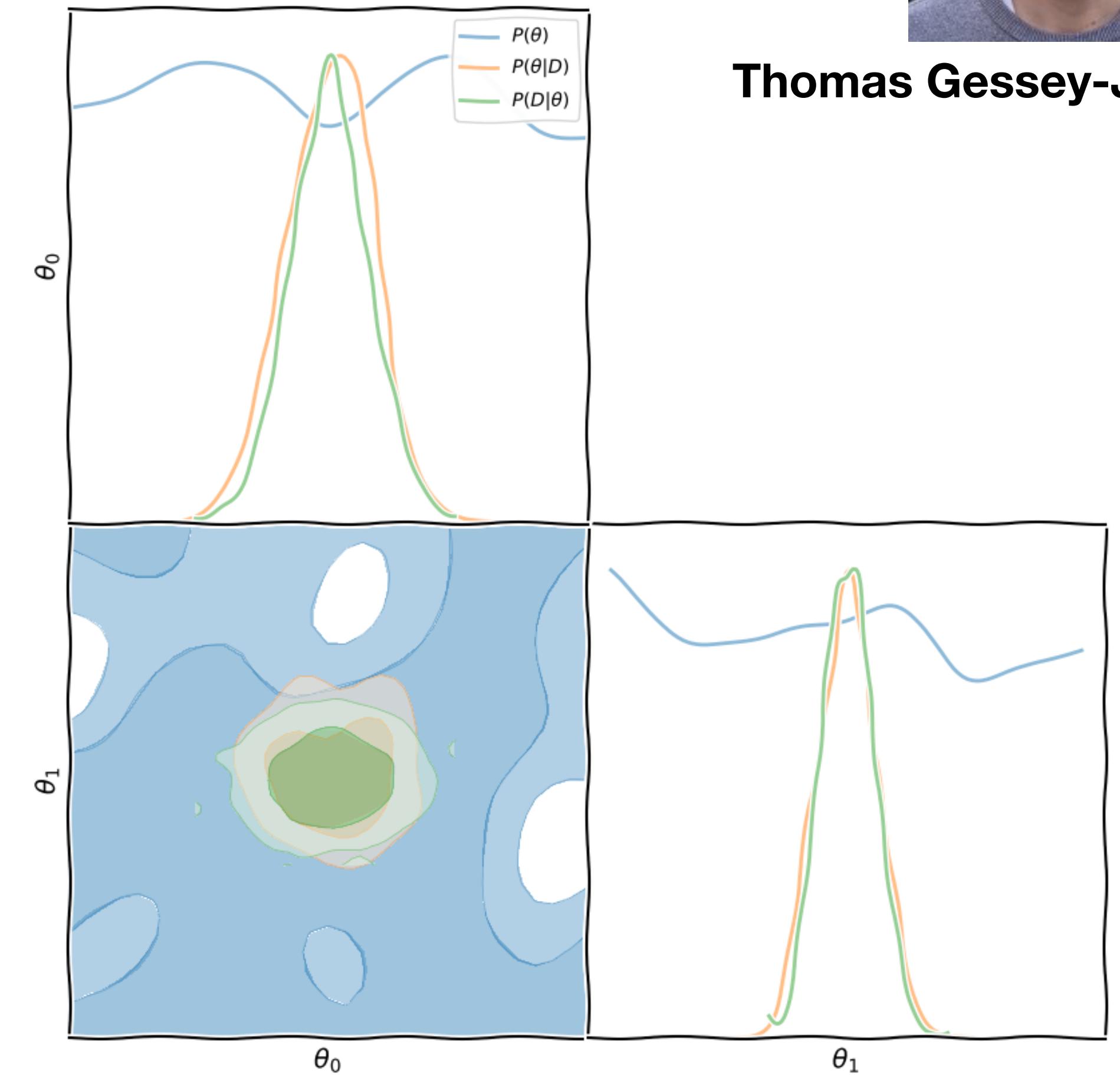
$$\pi = \mathcal{U}(\theta)$$

$$\rightarrow P = \mathcal{N}(\theta; C, \mu)$$

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

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

- Where C , C_E , μ and μ_E are functions of M , M_E and Σ the noise in the data



Measuring the impact of the emulator

- Assume a linear model and linear emulator error

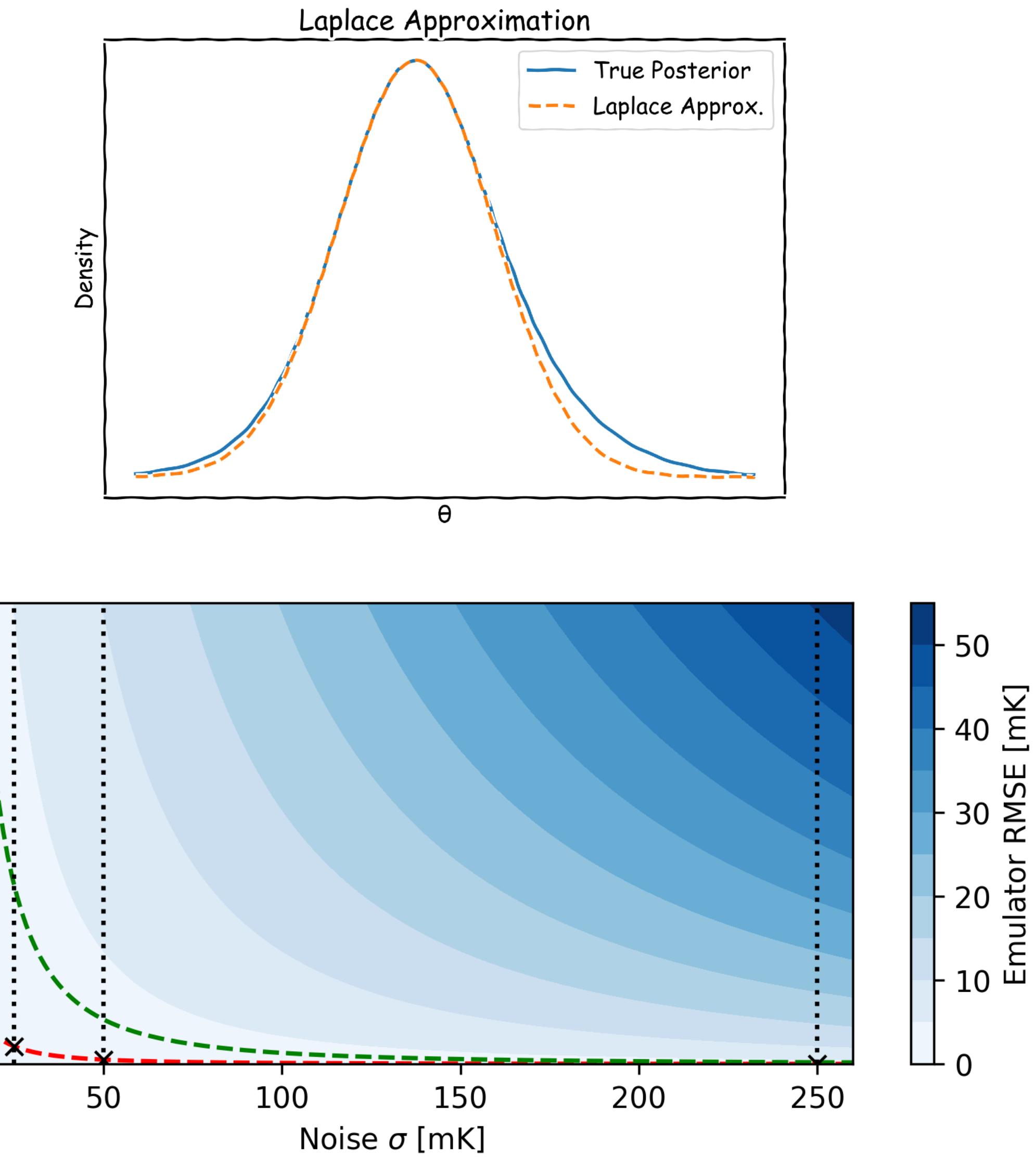
$$\mathcal{M}(\theta) \approx M\theta + m \text{ and } E(\theta) \approx E\theta + \epsilon$$

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

- Can show that

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

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

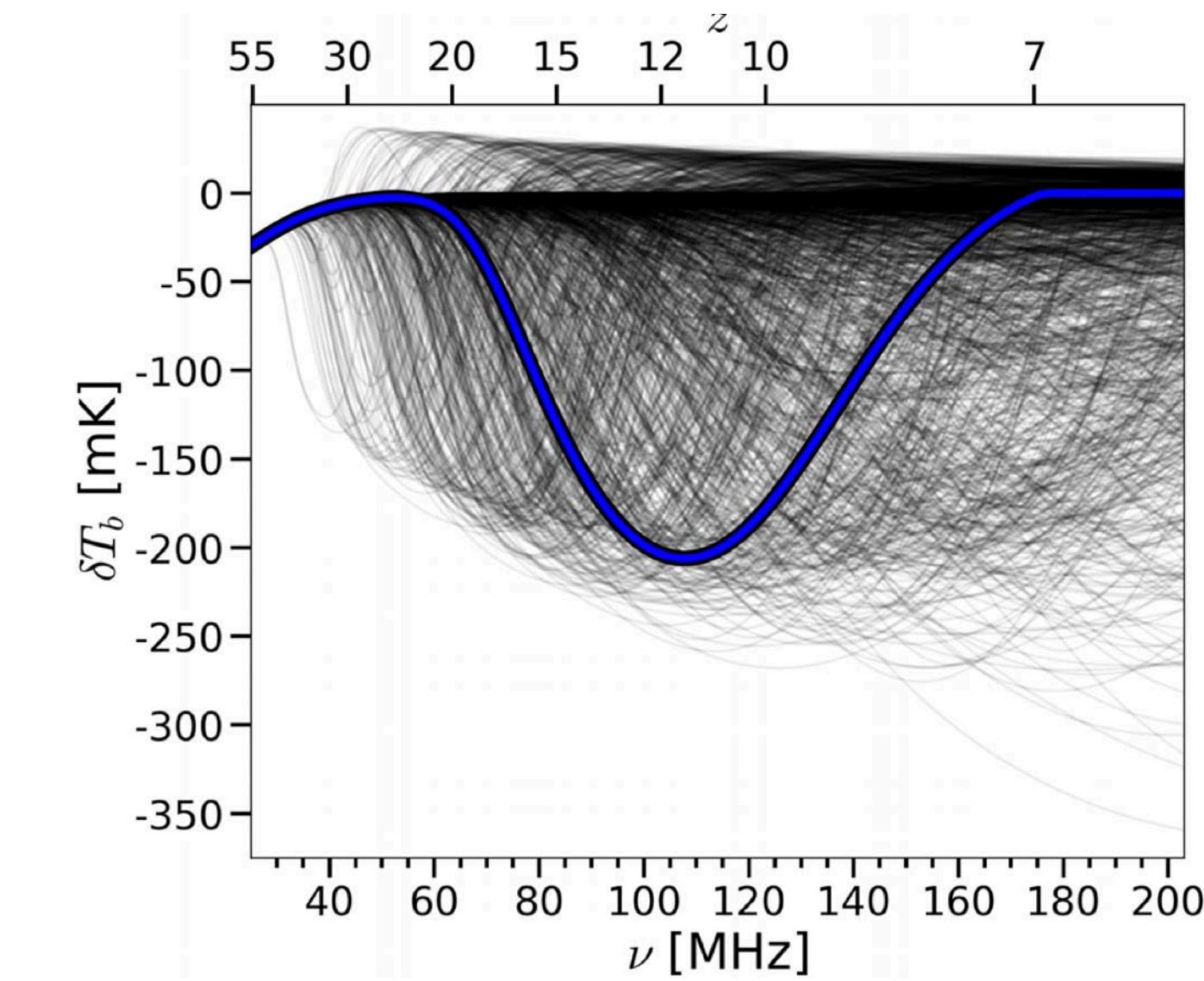
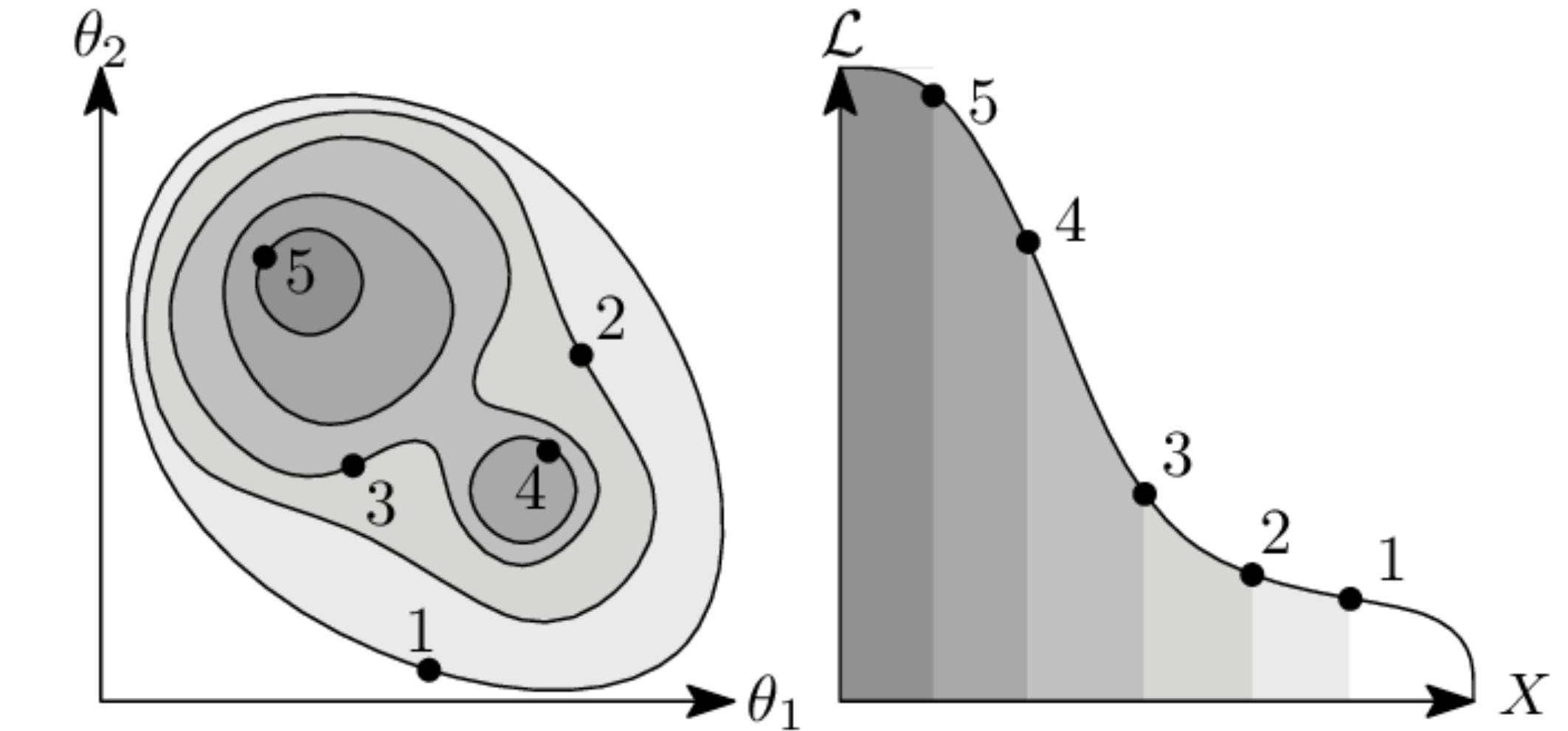


Measuring the impact of the emulator

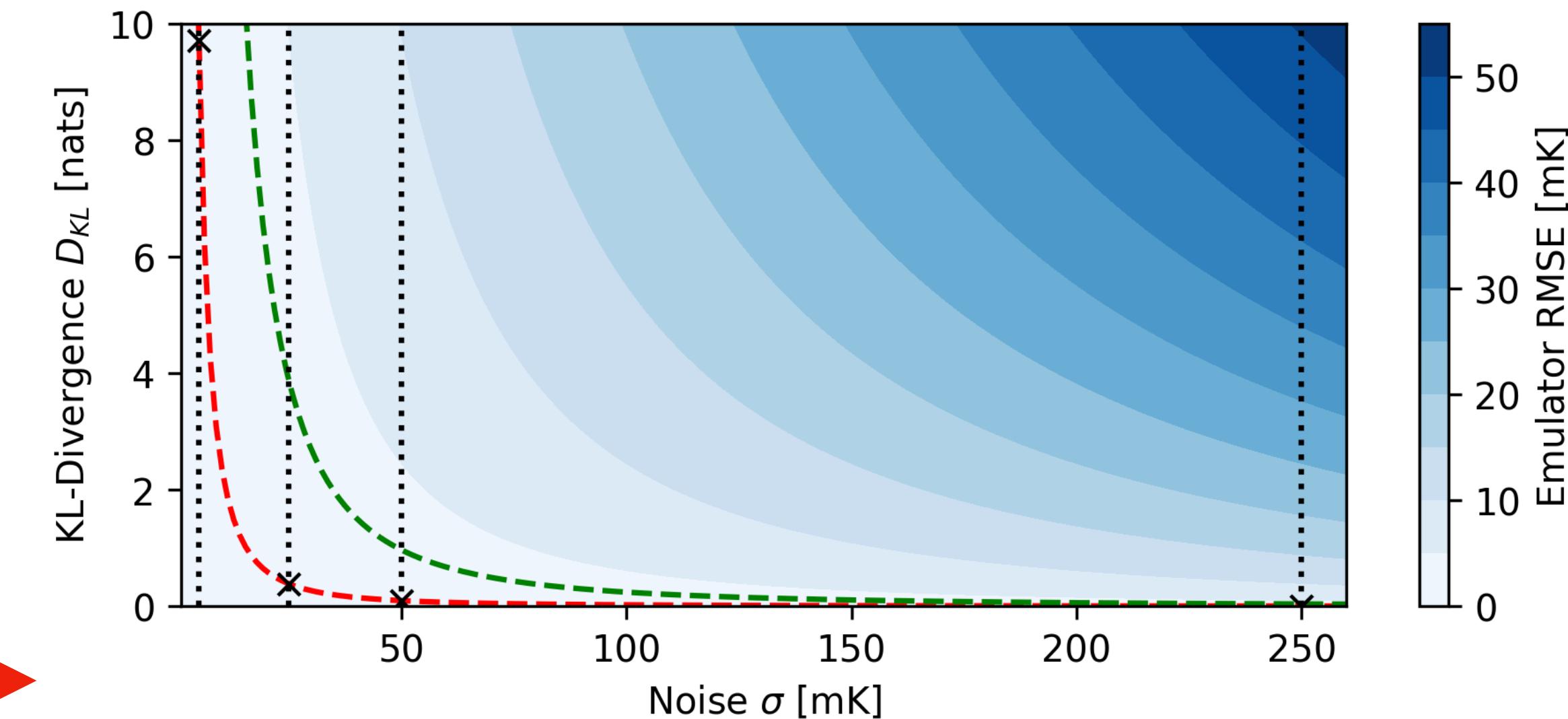
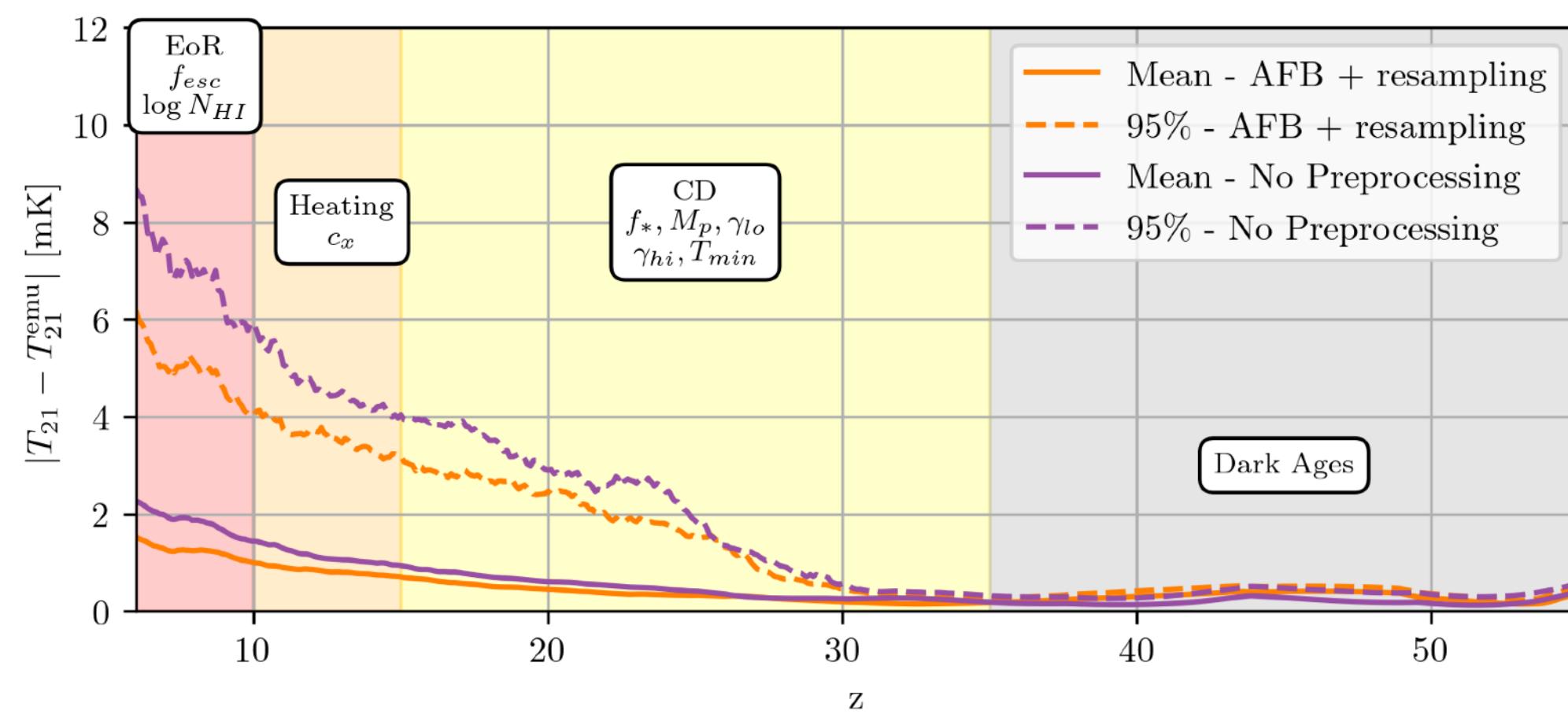
$$\epsilon = \sqrt{\frac{1}{N_\nu} \sum_i^{N_t} (S_{\text{true}}(t) - S_{\text{pred}}(t))^2} \quad \xrightarrow{\text{red arrow}} D_{\text{KL}}(P || P_E) \leq \frac{N_d}{2} \left(\frac{\text{RMSE}}{\sigma} \right)^2$$

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

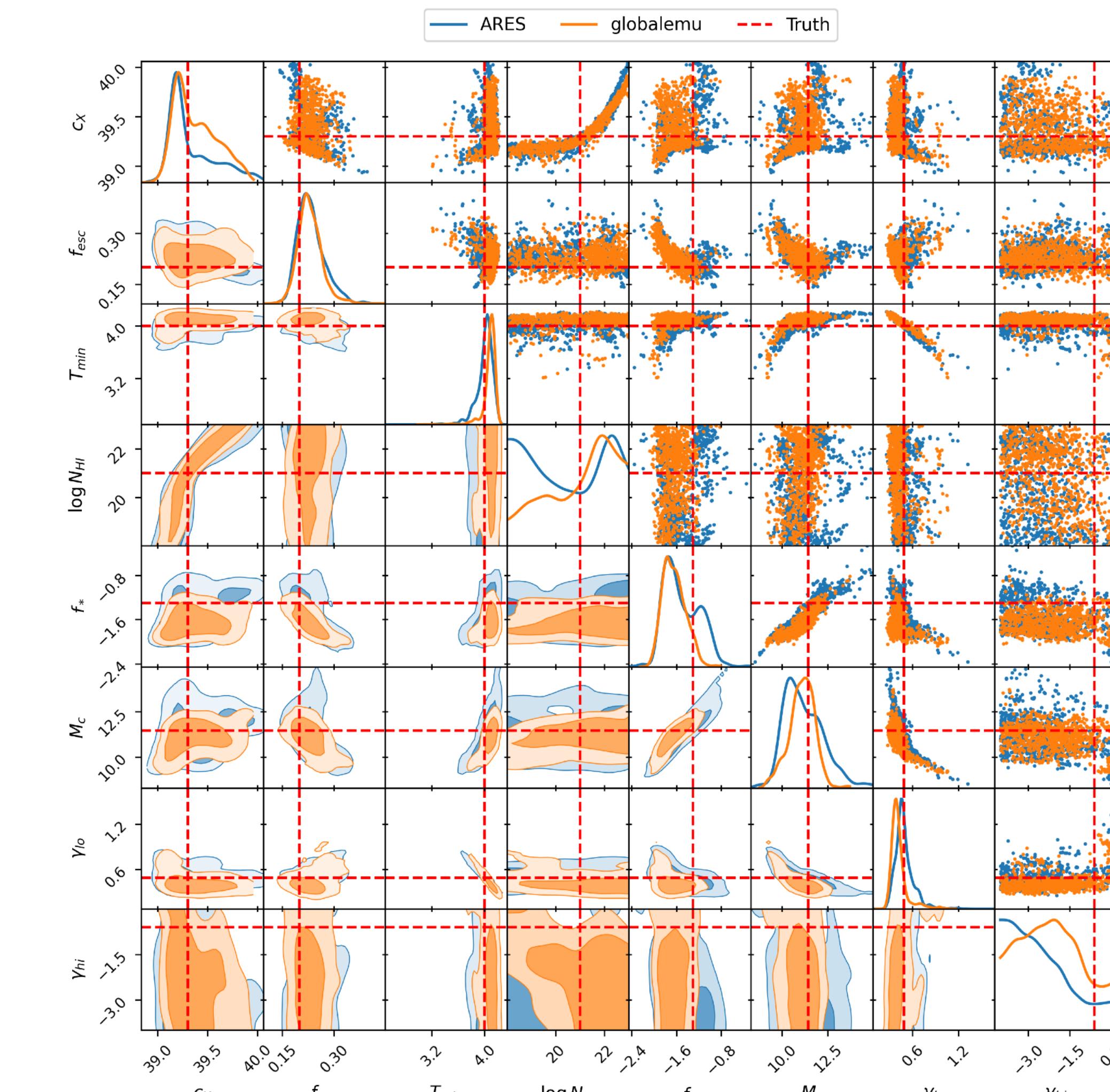
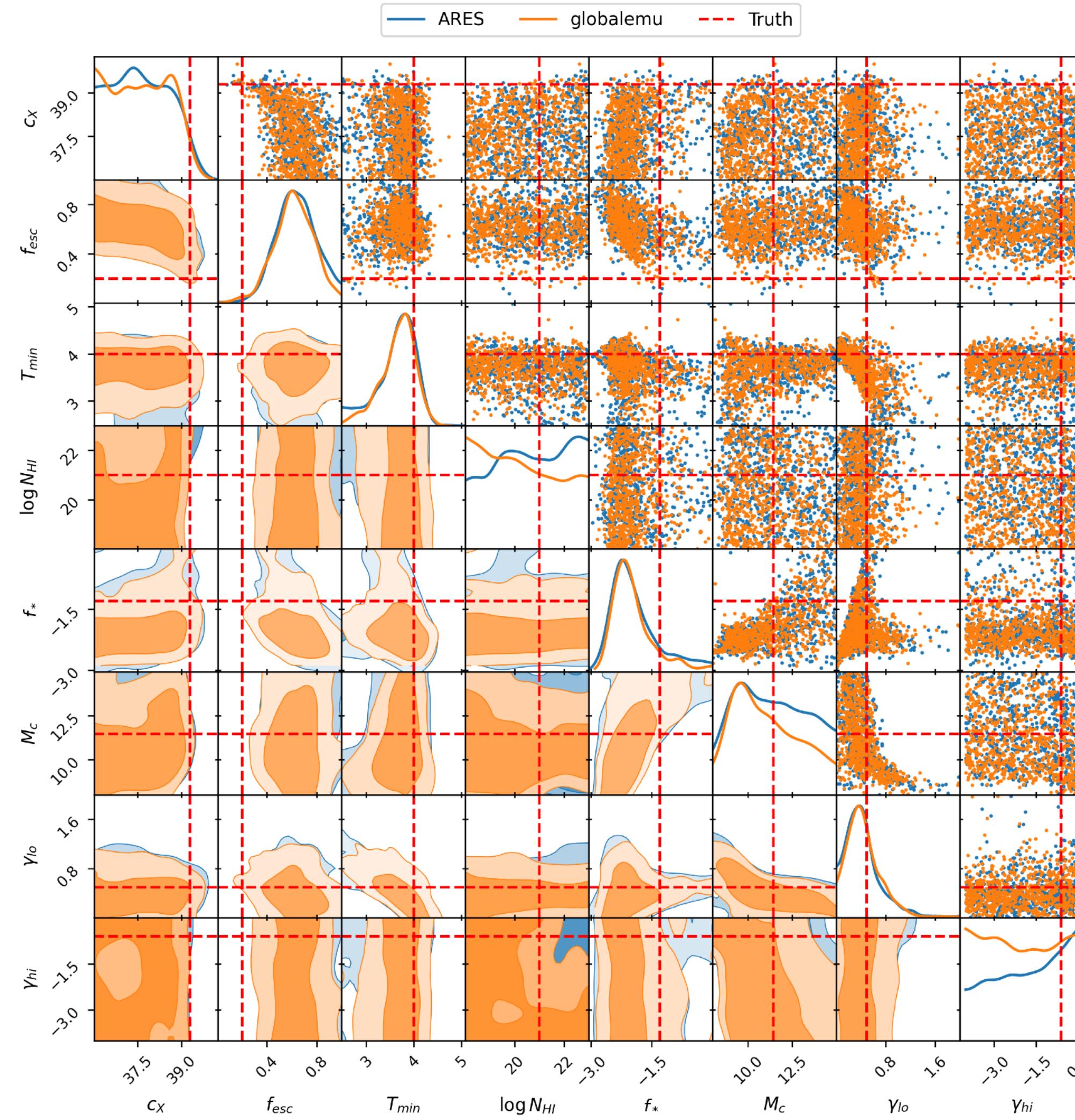


globalemu performance and ARES modelling

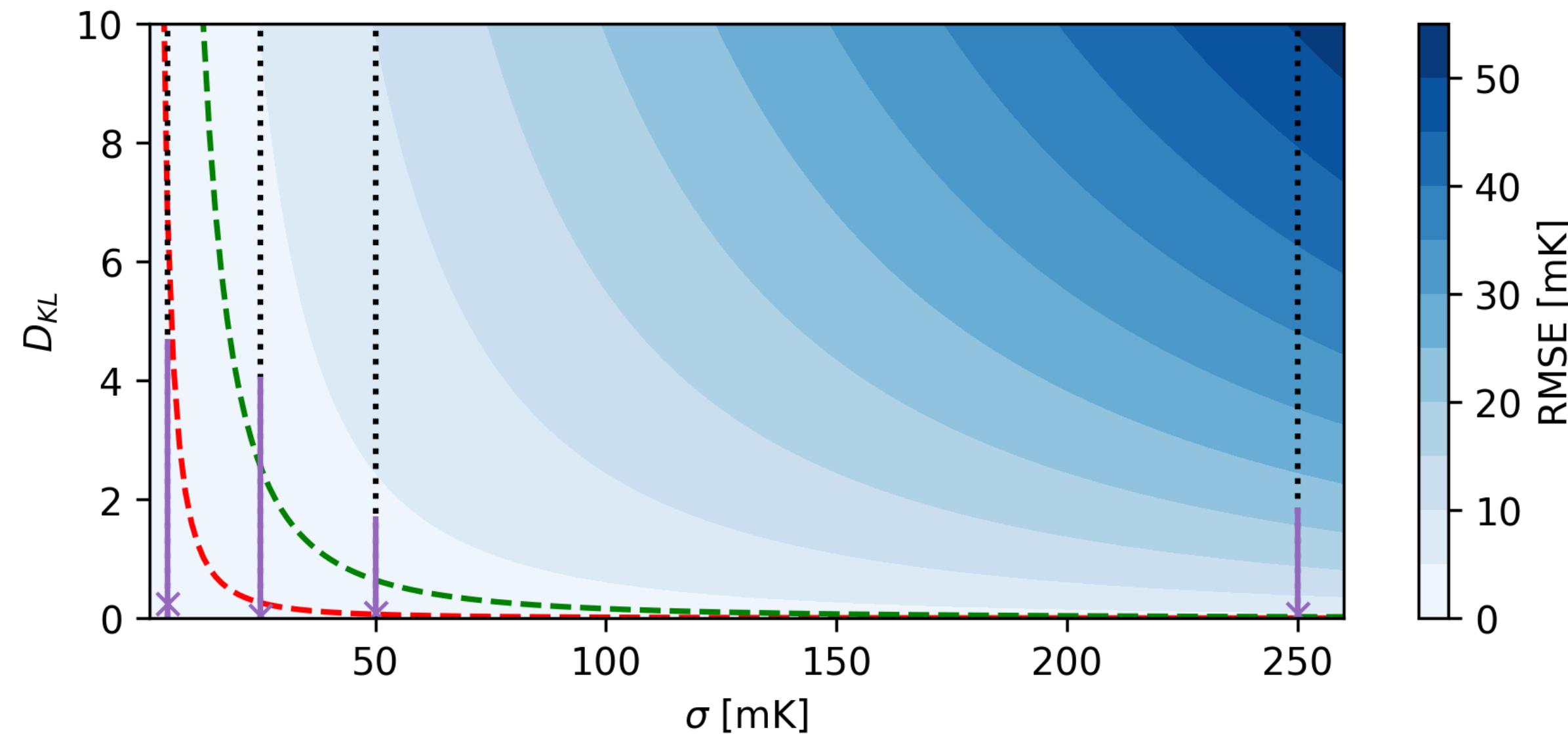


Noise Level [mK]	Estimated $\mathcal{D}_{KL} \leq$	
	Mean RMSE	95th Percentile
5	9.60	96.62
25	0.38	3.86
50	0.10	0.97
250	0.004	0.039

Running the analysis - 50 mK and 5 mK



How about the D_{KL} ?

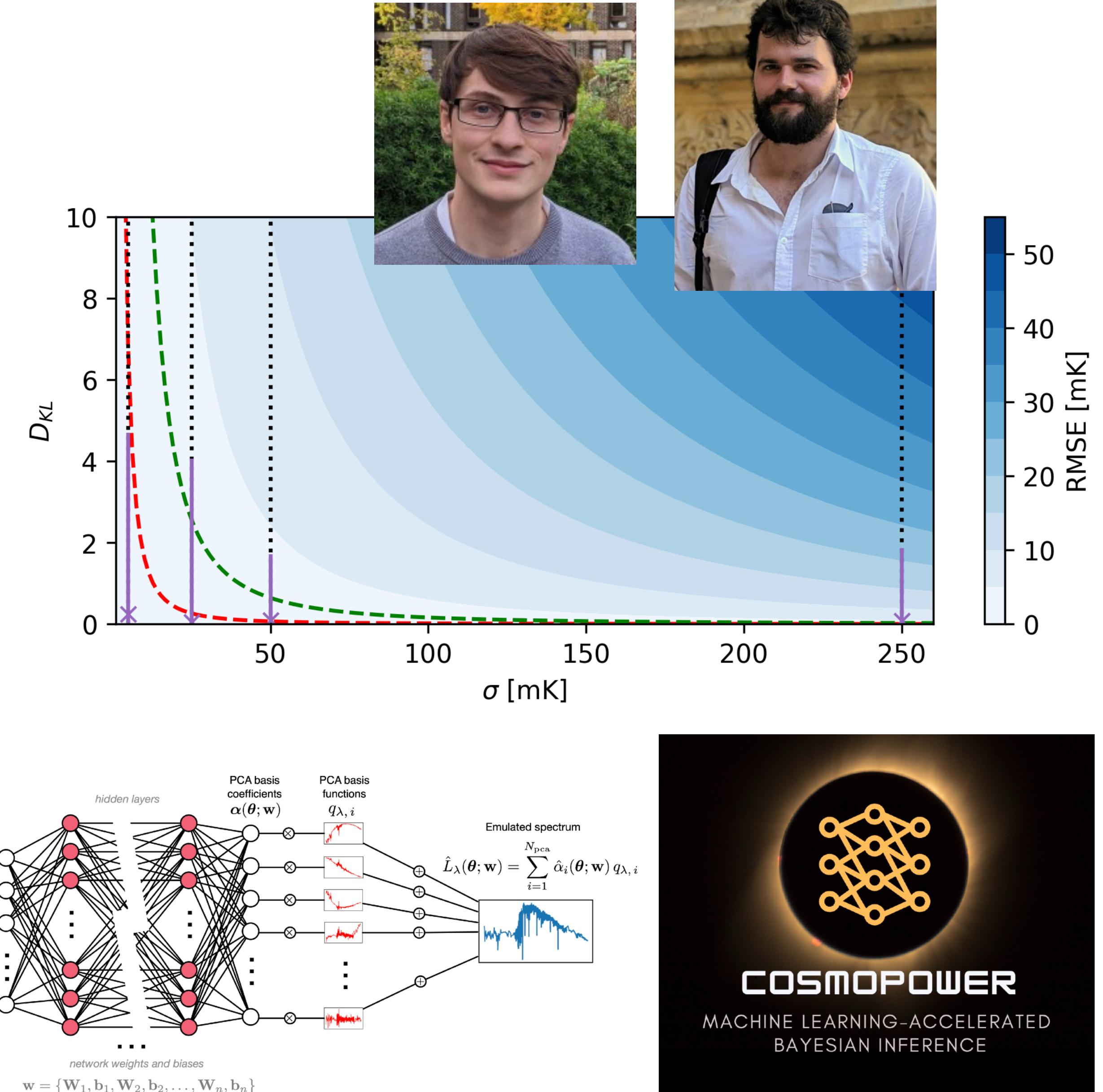


Noise Level [mK]	Estimated $\mathcal{D}_{KL} \leq$		Actual \mathcal{D}_{KL}
	Mean RMSE	95th Percentile	
5	9.60	96.62	$0.25^{+4.45}_{-0.25}$
25	0.38	3.86	$0.05^{+4.02}_{-0.52}$
50	0.10	0.97	$0.09^{+1.62}_{-0.03}$
250	0.004	0.039	$0.08^{+1.78}_{-0.02}$

- Use normalising flows implemented with *margarine* [see Bevins et al 2022, 2023, arXiv:2207.11457, arXiv:2205.12841]

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



Additional Slides

Measuring the impact of the emulator

- Assume a linear model and linear emulator error

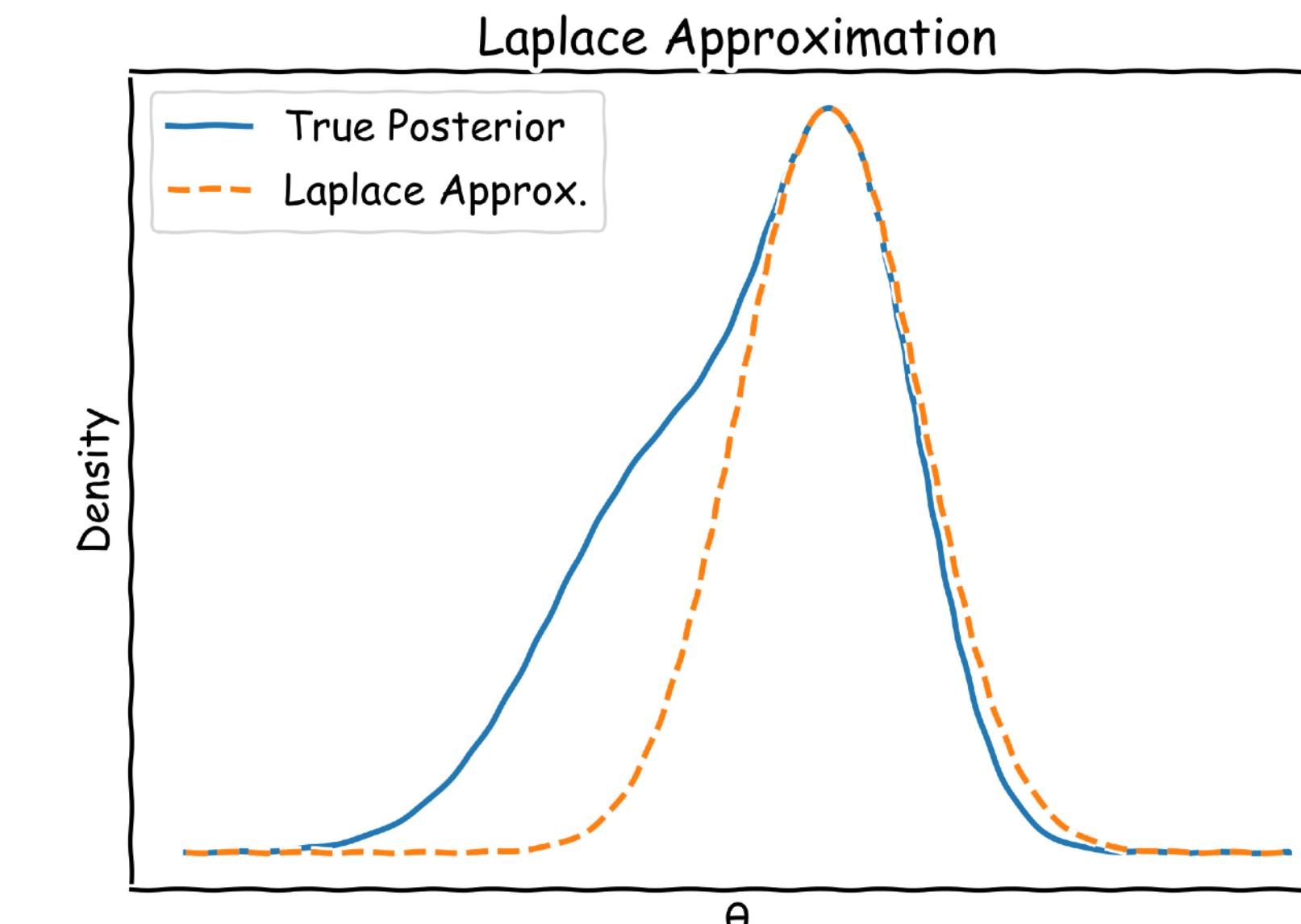
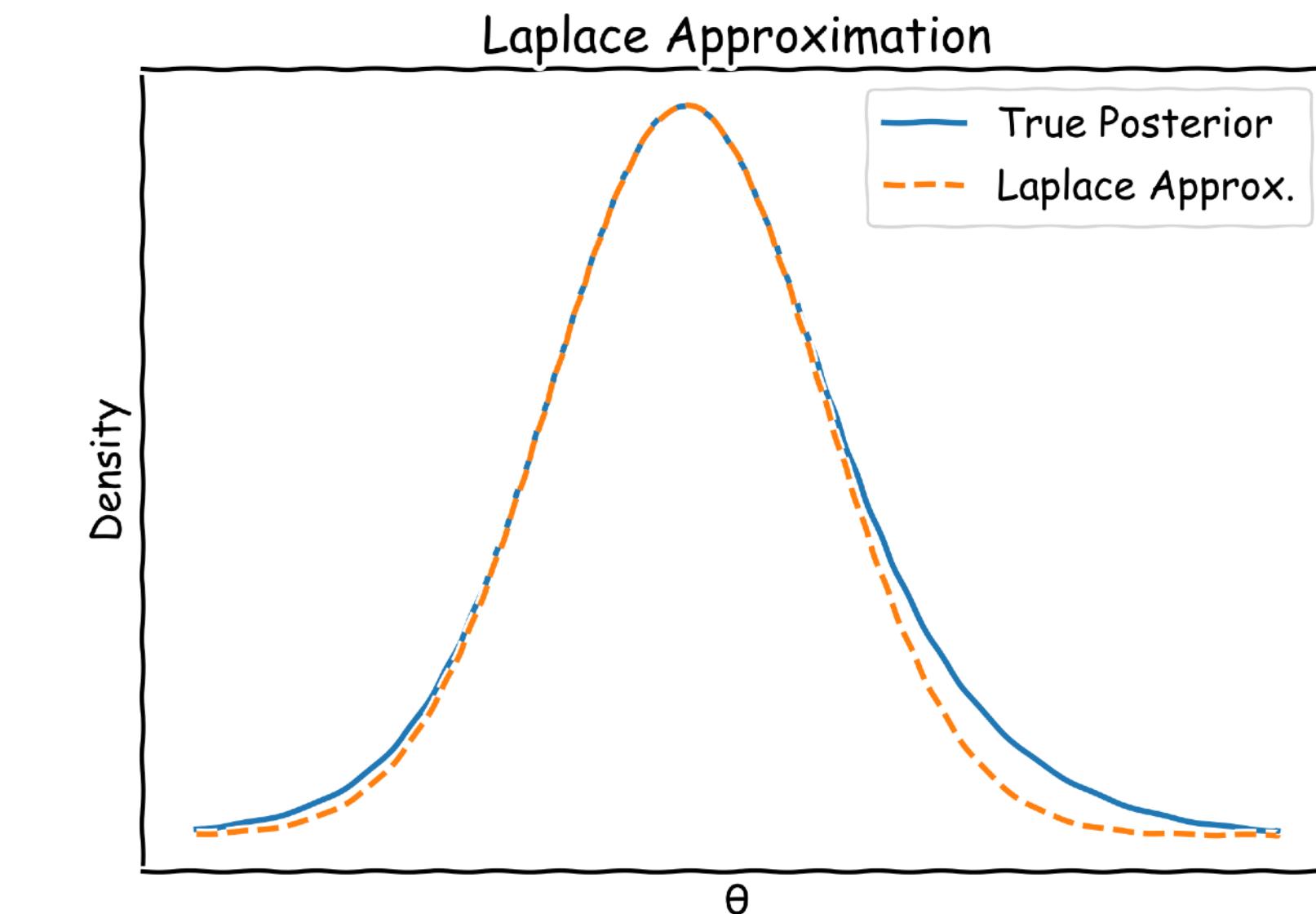
$$\mathcal{M}(\theta) \approx M\theta + m \text{ and } E(\theta) \approx E\theta + \epsilon$$

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 so we can ignore higher order terms

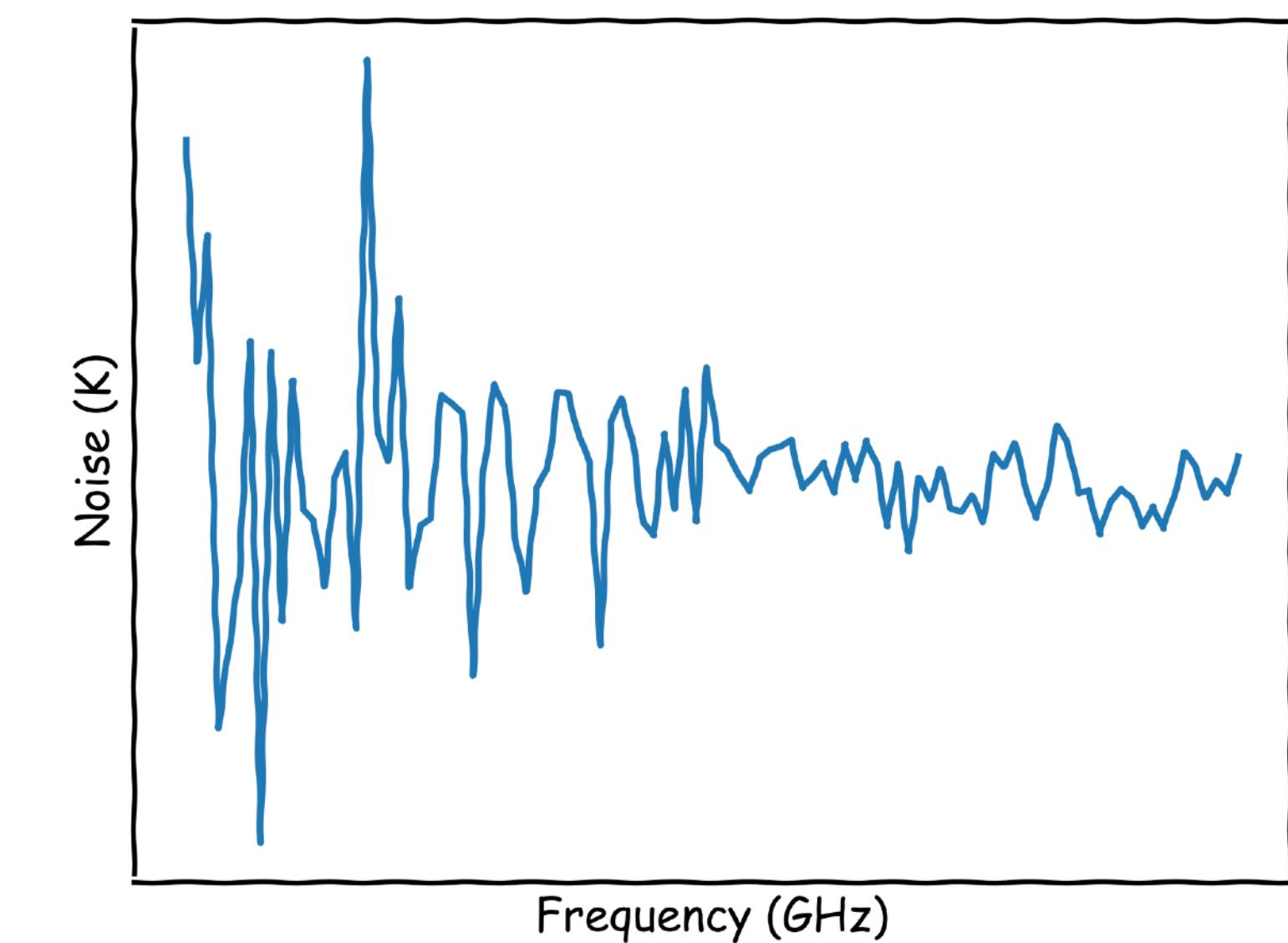
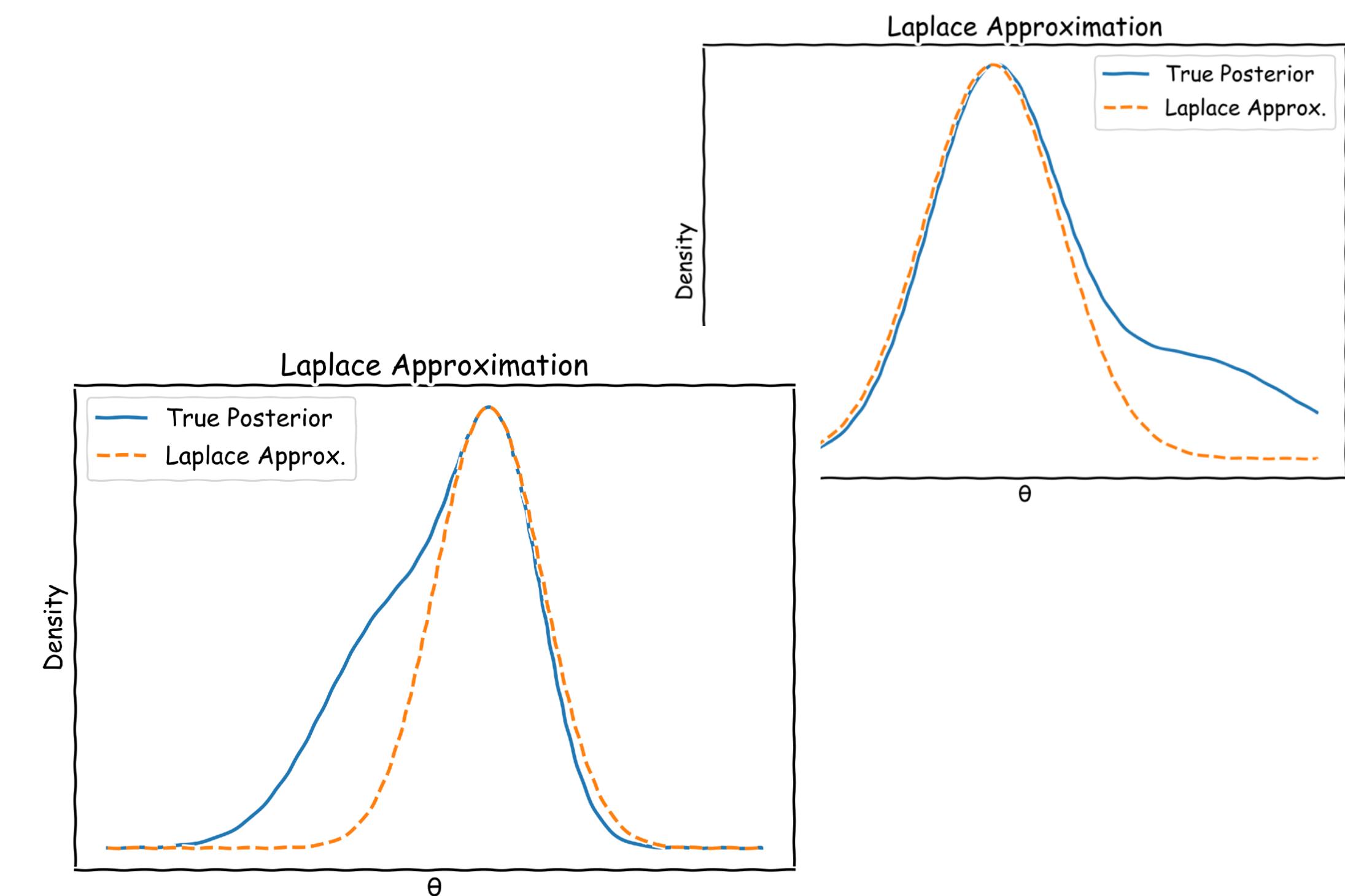
$$M = \mathcal{J}(\theta_0)$$

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

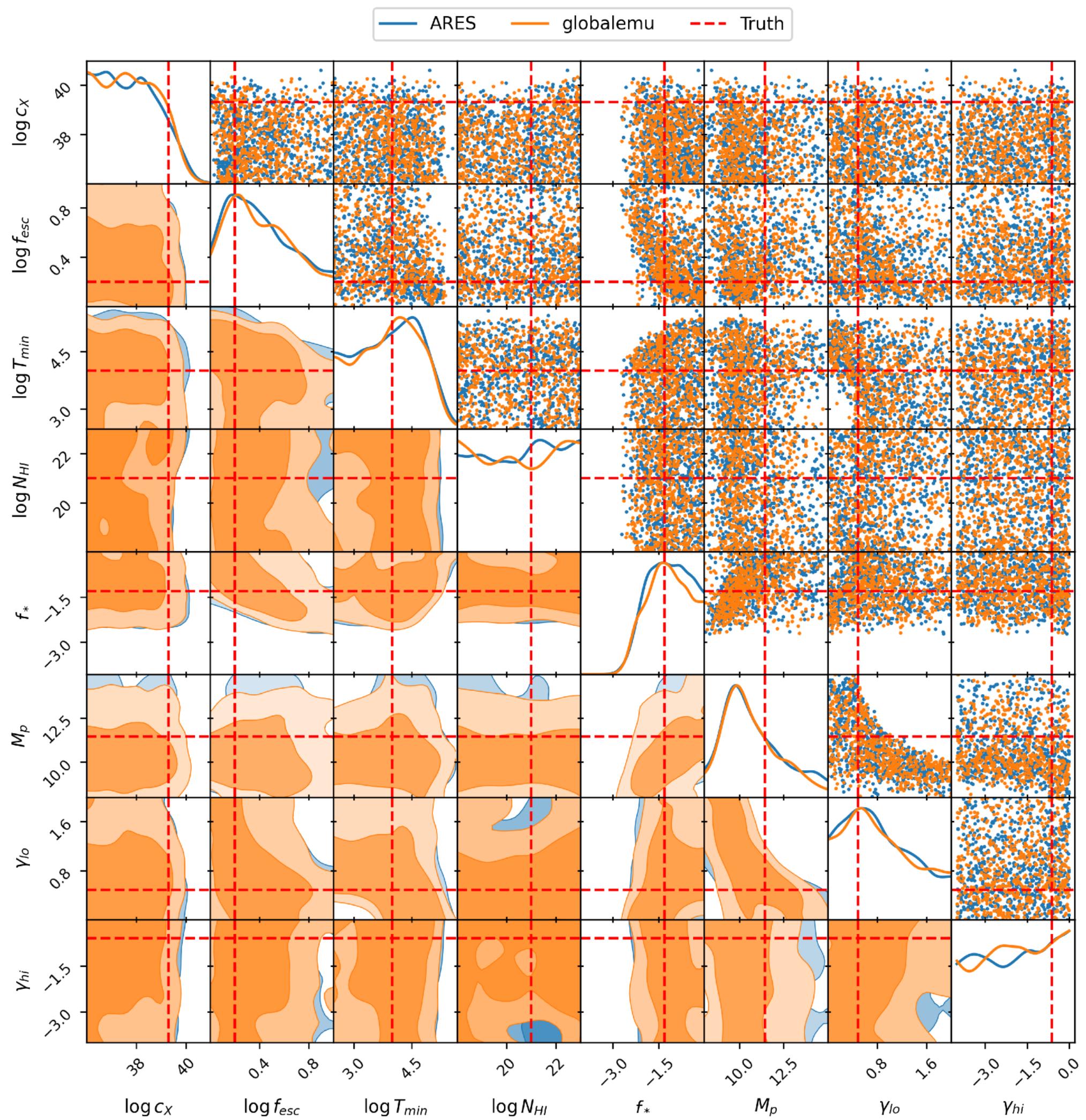


Limitations of the approximation

- 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



Running the analysis - 250 mK



Running the analysis - 25 mK

