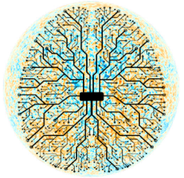
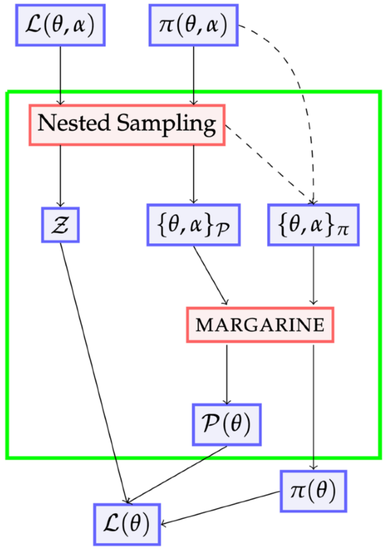
Machine Learning enhanced Bayesian Inference for Cosmology

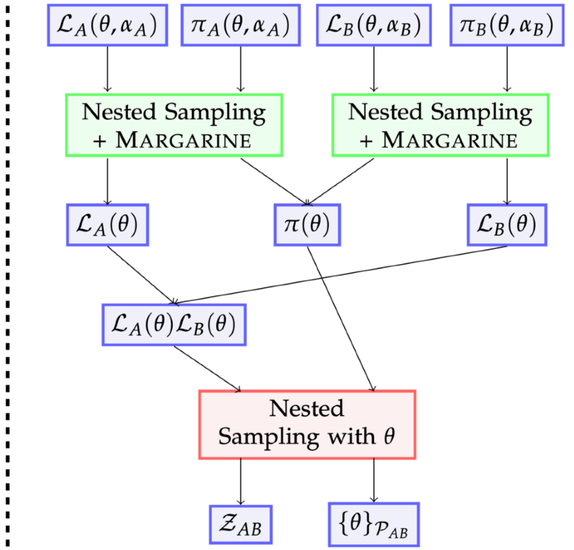
Harry Bevins


Contents

- I will discuss
  - Bayesian inference recap
  - Marginal Bayesian inference [2207.11457, 2205.12841]
  - Marginal Bayesian Statistics
  - Joint Analysis and Hierarchical Modelling





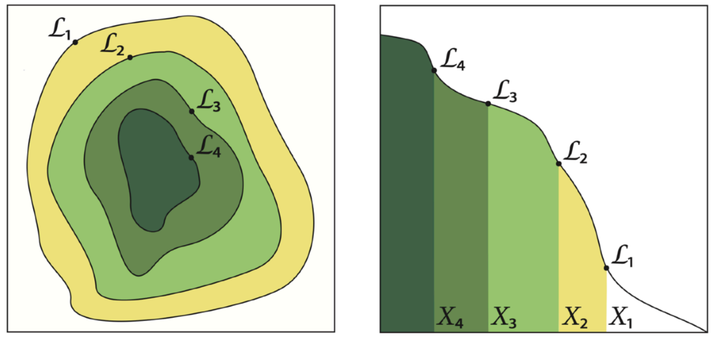
Bayesian Inference recap

Bayesian Inference

$$P(\Theta \mid D, M) = \frac{P(D \mid \Theta, M) P(\Theta \mid M)}{P(D \mid M)} = \frac{L(\Theta)\pi(\Theta)}{Z}$$

- Posterior $P(\Theta \mid D, M)$
- Likelihood $L(\Theta) = P(D \mid \Theta, M)$
- Prior $\pi(\Theta) = P(\Theta \mid M)$
- Evidence $Z = P(D \mid M)$
- Lots of different ways to access the posterior and evidence (e.g. Metropolis Hastings, HMC, SMC, Nested Sampling, ML-enhanced Harmonic Means, *floZ,* Simulation Based Inference)
- Focusing on Nested Sampling here

emcee
The MCMC Hammer

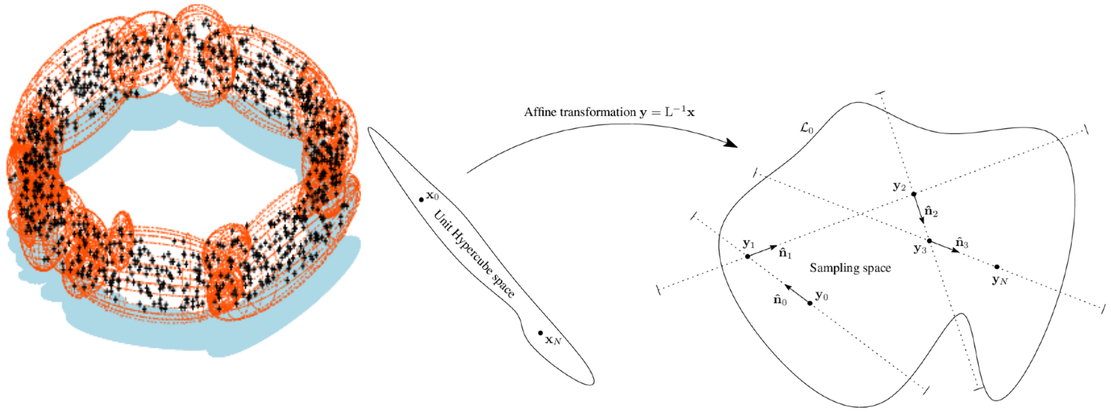


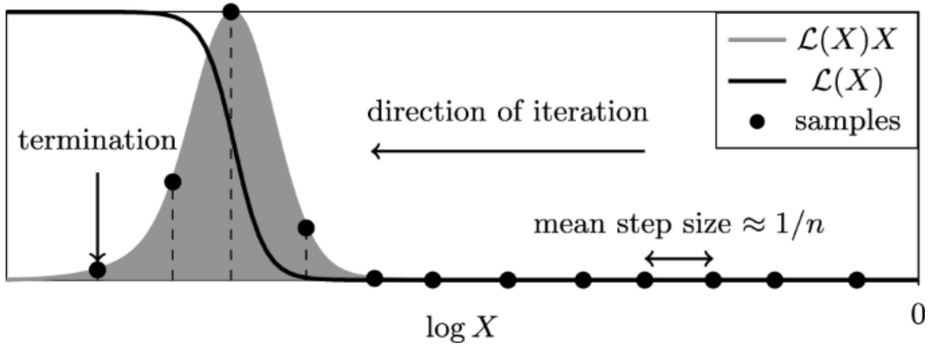
sbi

Scaling of Nested Sampling

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

- n_{live} is number of explorers
- $\langle T\{L(\Theta)\}\rangle$ is the time complexity of the likelihood
- $\langle T\{Impl.\}\rangle$ is time complexity of sampling method e.g. slice sampling, region sampling etc
- $D_{KL}(P\,||\,\pi)$ is the KL divergence between posterior and prior

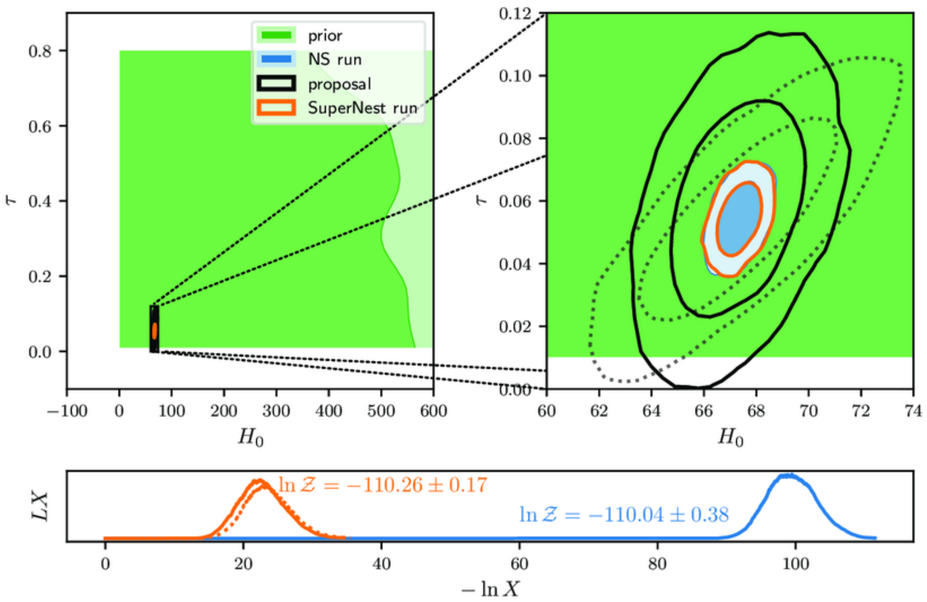




Scaling of Nested Sampling

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

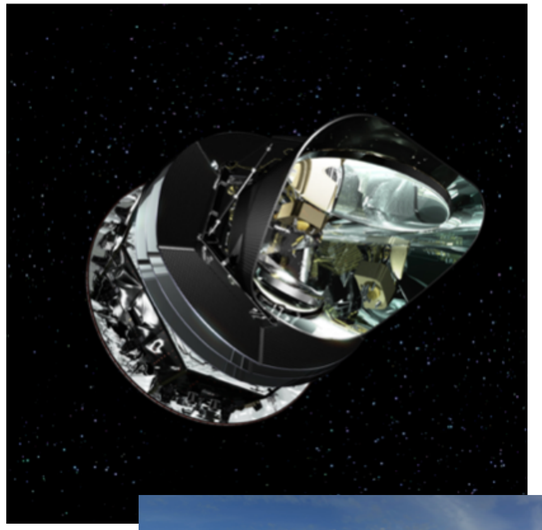
- Have to be careful fiddling with n_{live} (dynamic nested sampling)
- Improving $\langle T\{\text{Impl.}\}\rangle$ is hard!
- Reducing the KL-divergence is eminently doable (see 2212.01760)
- Speeding up our likelihood is also usually doable
- For Nested Sampling it can be shown that $T \propto d^3$

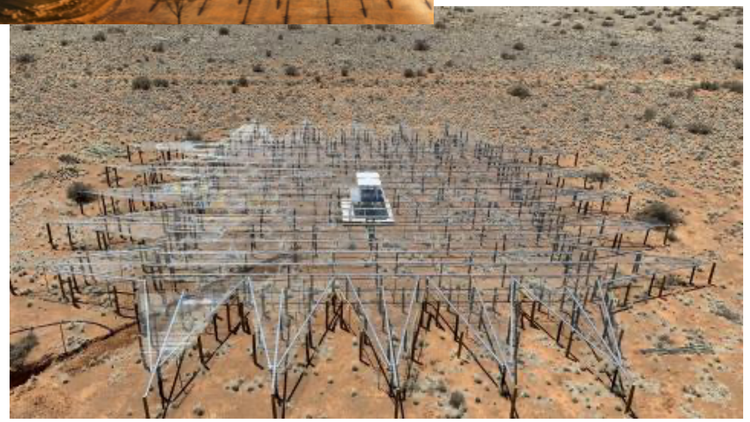


Marginal Bayesian Inference

Why are we interested in the marginal space?

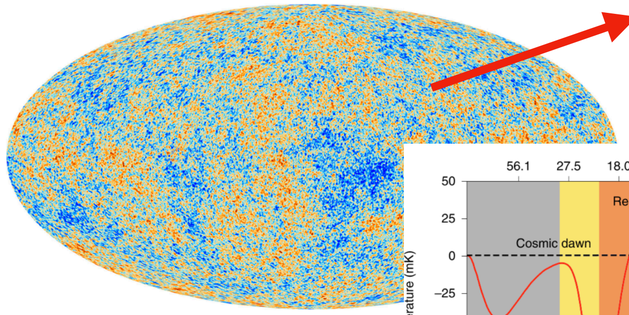
- Often we have nuisance parameters α in our modelling that describe instrumental effects or contaminating signals
- While they are interesting we usually are only *really* interested in a few cosmological parameters θ
- Nuisance parameters make joint analysis hard and make comparing experimental results difficult

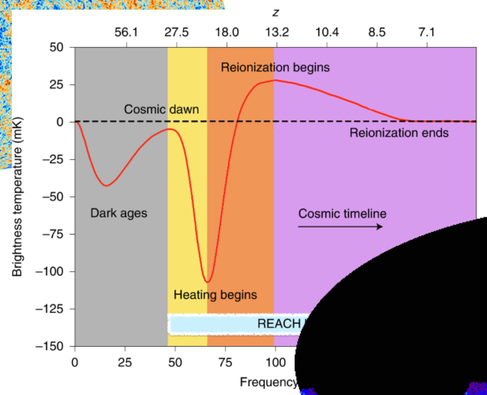




The 21-cm Line

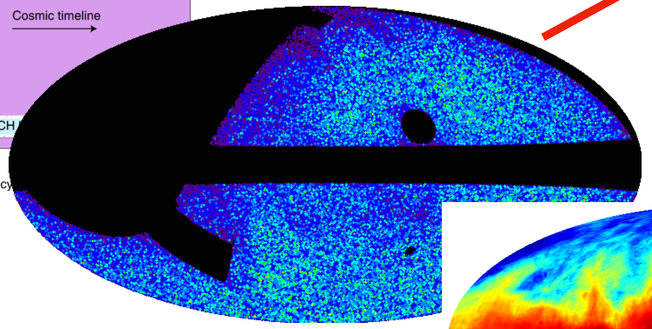
The Cosmic Microwave Background

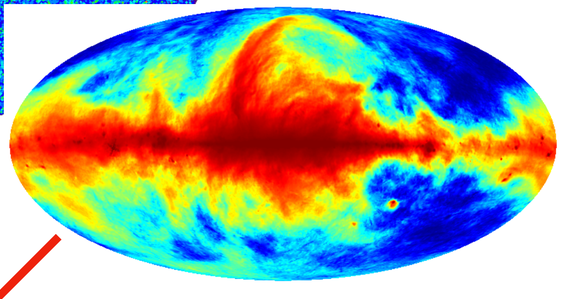




The 21-cm signal
(What we are interested in!)

Foreground galaxies

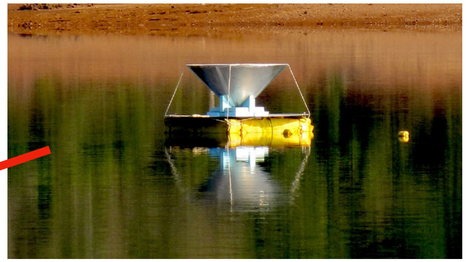




Our own Galaxy

The atmosphere

The environment around the telescope



The marginal likelihood

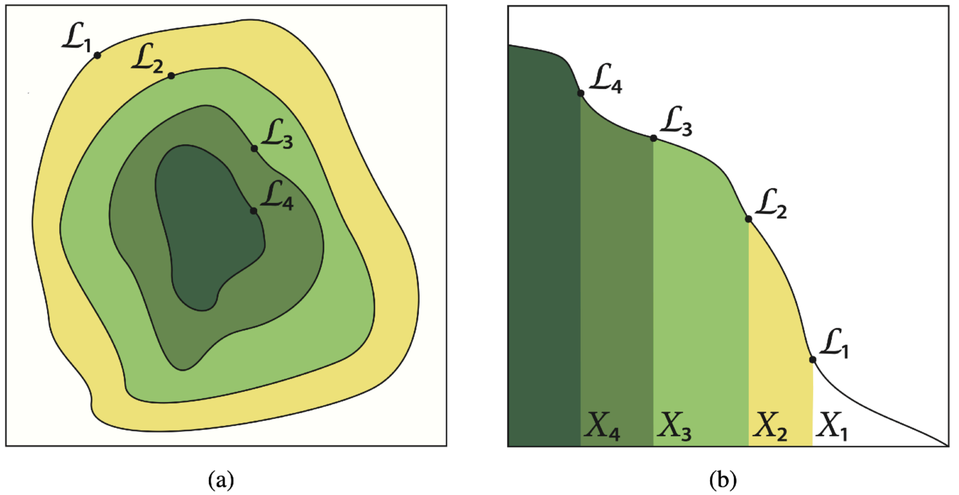
- If we define the marginal posterior and prior as

$$P(\theta) = \int P(\theta, \alpha) d\alpha \quad \text{and} \quad \pi(\theta) = \int \pi(\theta, \alpha) d\alpha$$

- Then we can define the nuisance marginalised likelihood as

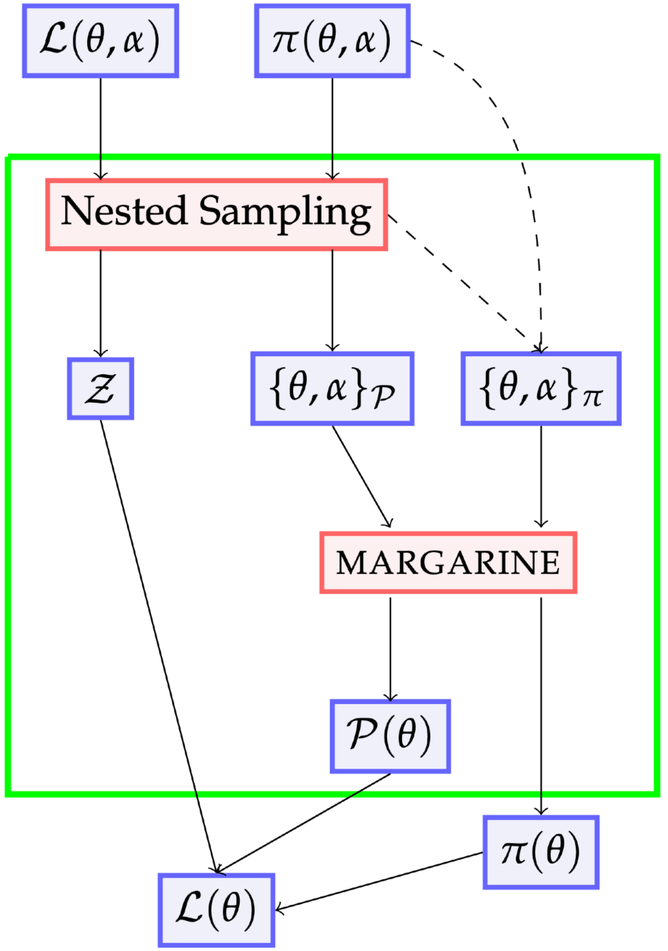
$$L(\theta) = \frac{\int L(\theta, \alpha)\pi(\theta, \alpha) d\alpha}{\int \pi(\theta, \alpha) d\alpha} = \frac{P(\theta)Z}{\pi(\theta)}$$

- Allows us to do efficient joint inference



Normalising Flows

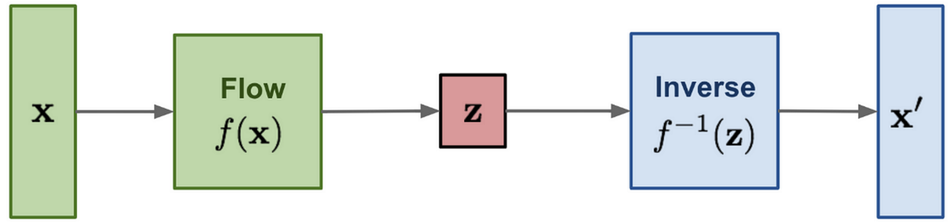
- The issue is that we have samples on $\Theta_P = \{\theta, \alpha\}_P \sim P(\theta, \alpha)$ and $\Theta_\pi = \{\theta, \alpha\}_\pi \sim \pi(\theta, \alpha)$ but not the marginal probabilities $P(\theta)$ and $\pi(\theta)$
- But we can access these with density estimation tools like Normalising Flows
- We implement Masked Autoregressive Flows and package our code up in to a python package called ***margarine***

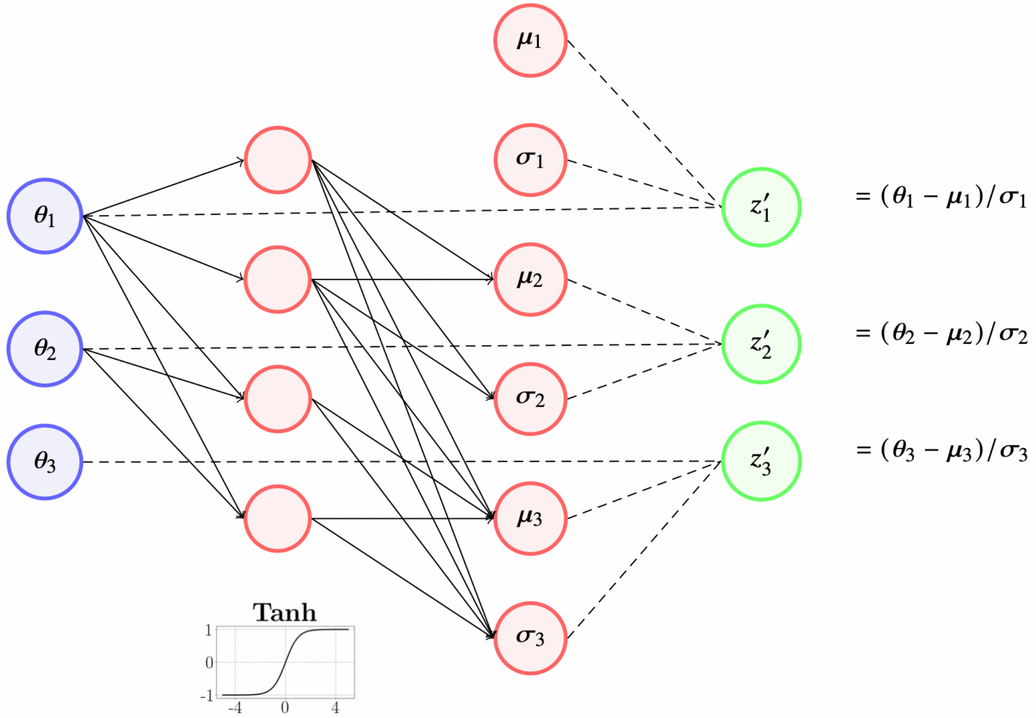


Normalising Flows

- Normalising flows are a class of machine learning model
- Learn a transformation from samples on a tractable distribution to samples on a more complex target like a posterior
- Invertible transformations

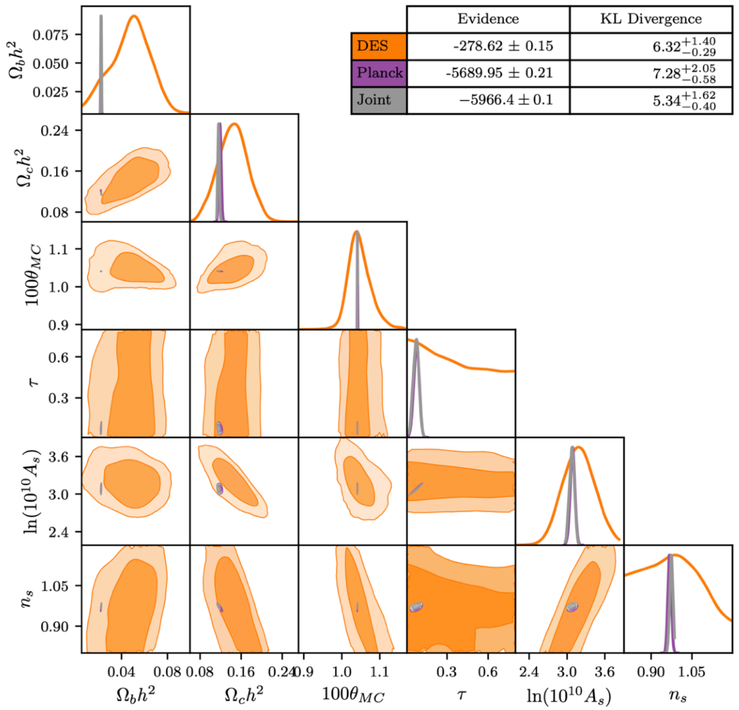
$$P(x) = P(f^{-1}(x)) \left| \det \left(\frac{df^{-1}(x)}{dx} \right) \right|$$





Why is this important?

- With these tools we can emulate $P(\theta)$, $L(\theta)$ and $\pi(\theta)$
- These emulators have a lot of applications
- Can find our code at: https://github.com/htjb/margarine
- And the relevant papers 2207.11457 and 2205.12841



margarine: Posterior Sampling and Marginal Bayesian Statistics

Introduction

margarine:	Marginal Bayesian Statistics
Authors:	Harry T.J. Bevins
Version:	1.2.8
Homepage:	https://github.com/htjb/margarine
Documentation:	https://margarine.readthedocs.io/

docs passing | launch binder | astro.IM arXiv:2205.12841

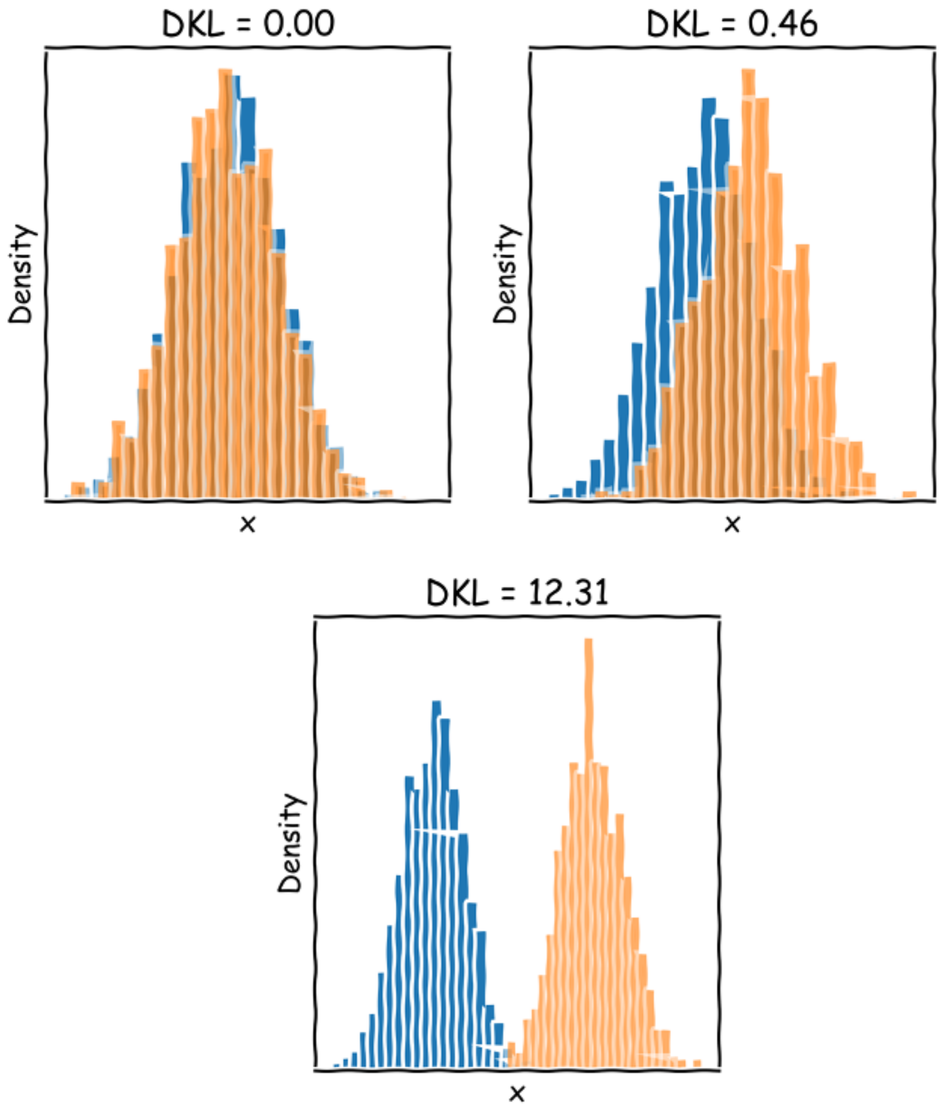
Marginal Bayesian Statistics

How do we estimate information gain?

- Typically we are interested in how informative an experiment is
- In a Bayesian sense this corresponds to the information gain between the prior and the posterior
- We can measure this with the Kullback-Leibler Divergence

$$D_{KL}(P\,||\,\pi) = \int P(\Theta)\log\frac{P(\Theta)}{\pi(\Theta)}d\Theta$$

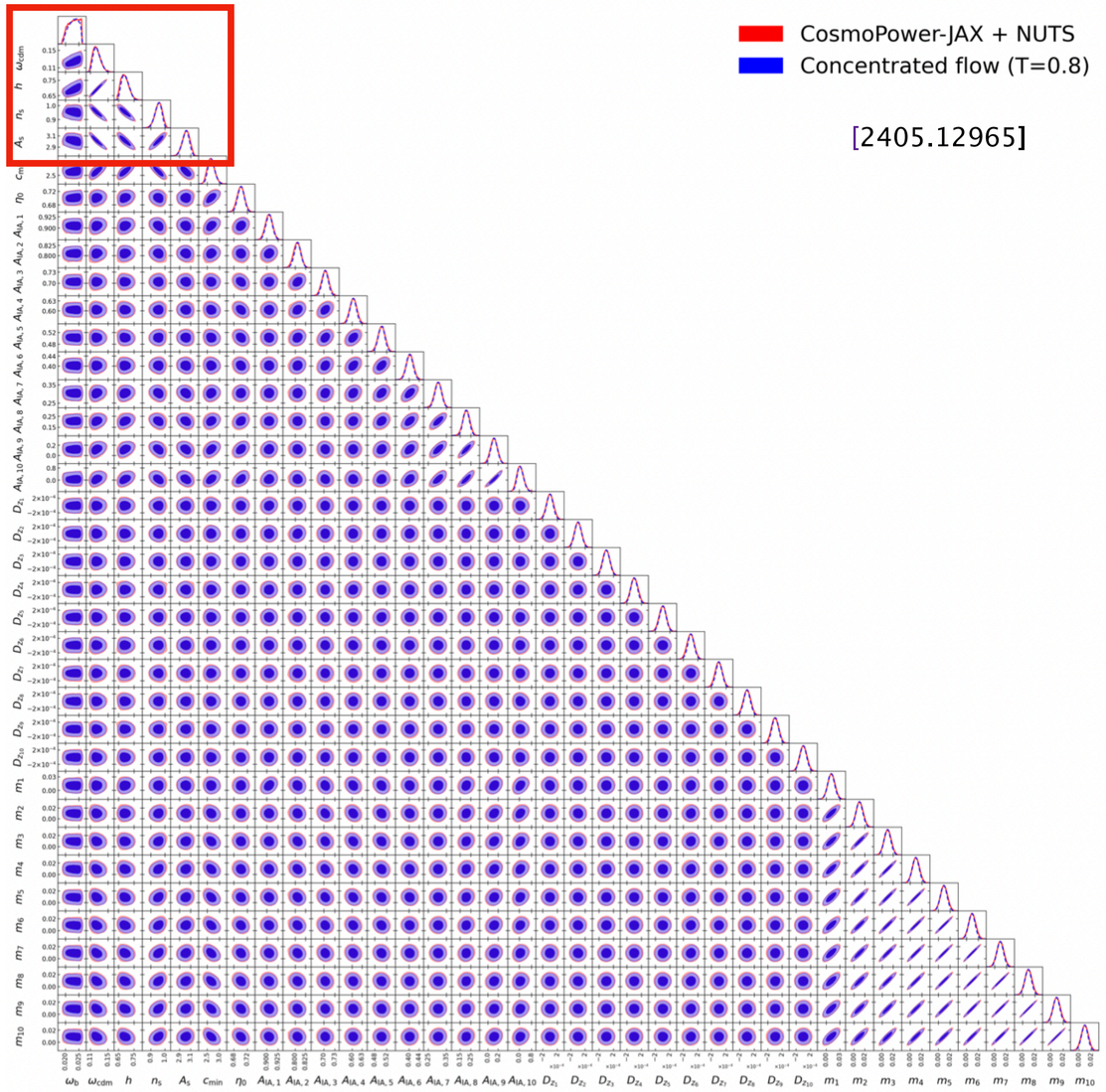
- And model dimensionalities



Comparing KL divergences

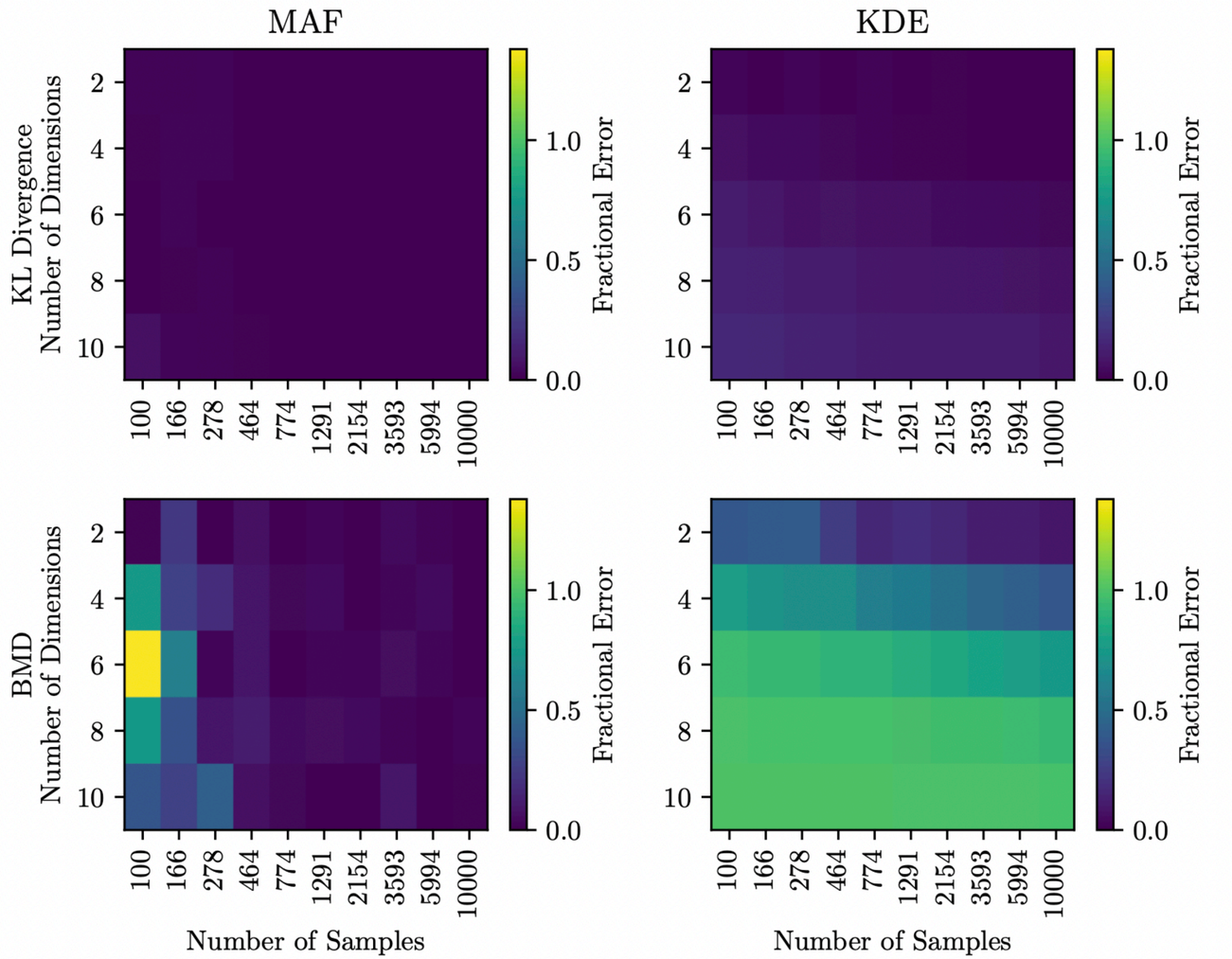
$$D_{KL}(P||\pi) = \int P(\Theta) \log \frac{P(\Theta)}{\pi(\Theta)} d\Theta$$

- What if we want to compare the information gain from different experiments?
- Well these experiments often have different sets of nuisance parameters
- Only really interested in the constraining power on the common parameters



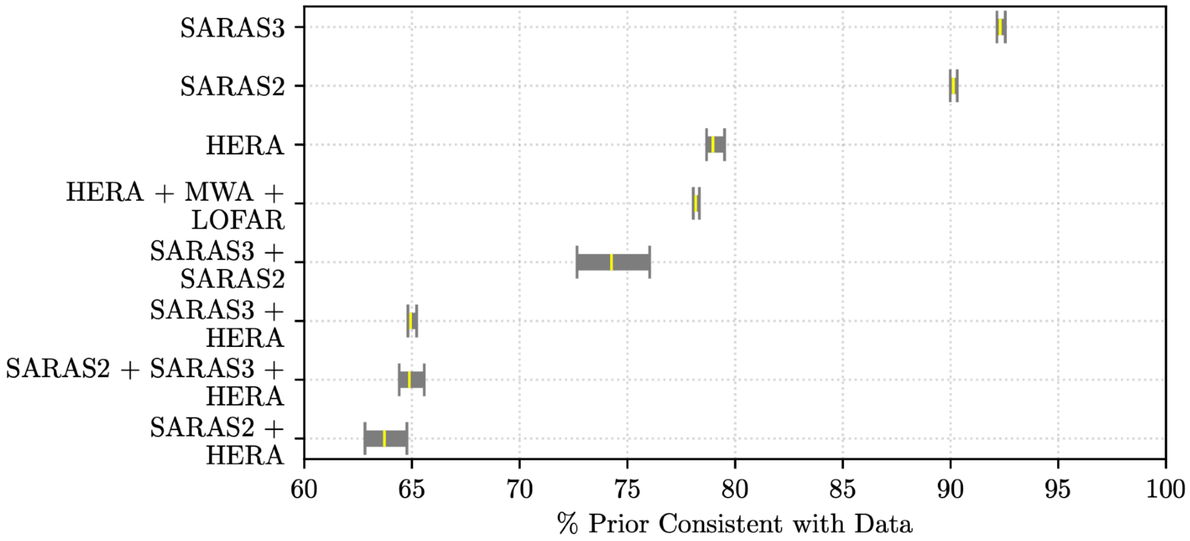
Accuracy of calculation

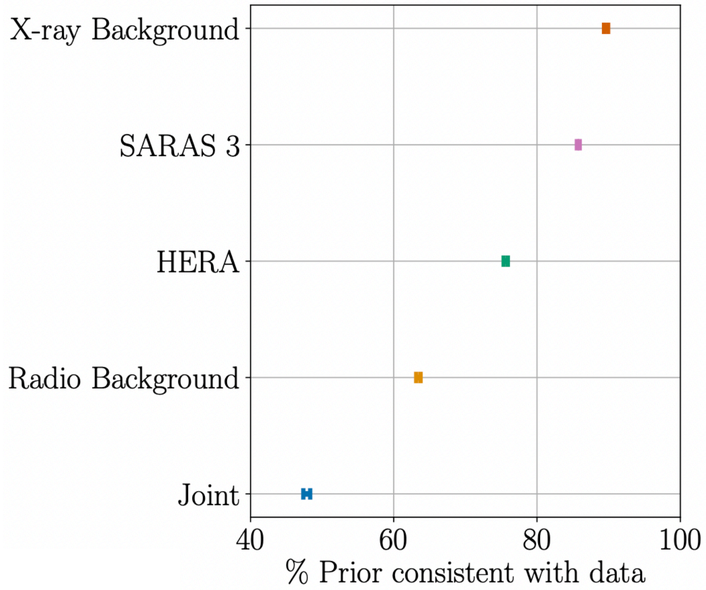
- In practice we can replace “normalising flows” with any density estimation tool
- In 2207.11457 we compared the accuracy of the KL divergence estimates on a known problem with KDEs and a class of NFs
- Multivariate 5D Gaussian distribution
- Currently discussing using diffusion models



Why is this useful?

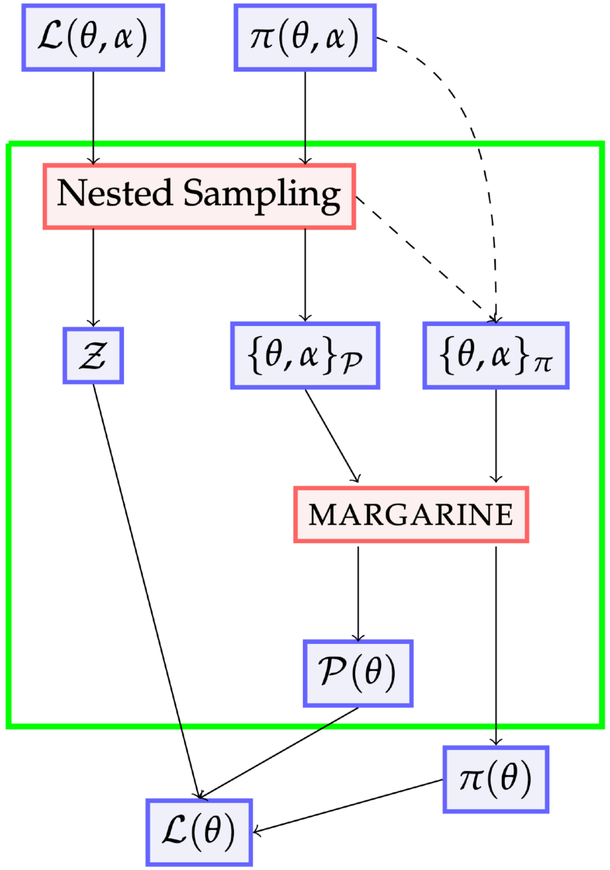
- Allows us to look at two different experimental approaches and say whether one is more powerful over another
- Use this to inform experimental design
- And as evidence for building the next generation of a particular class of experiments



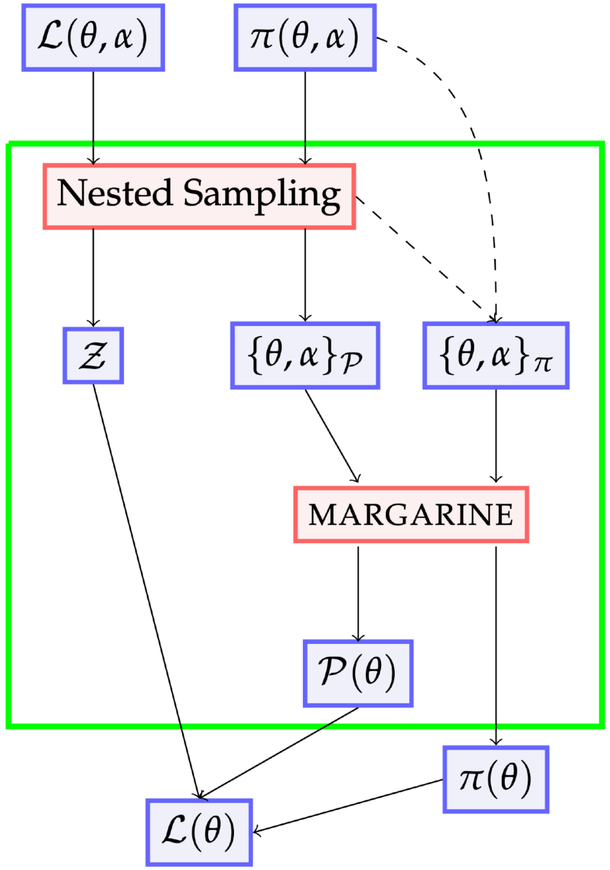


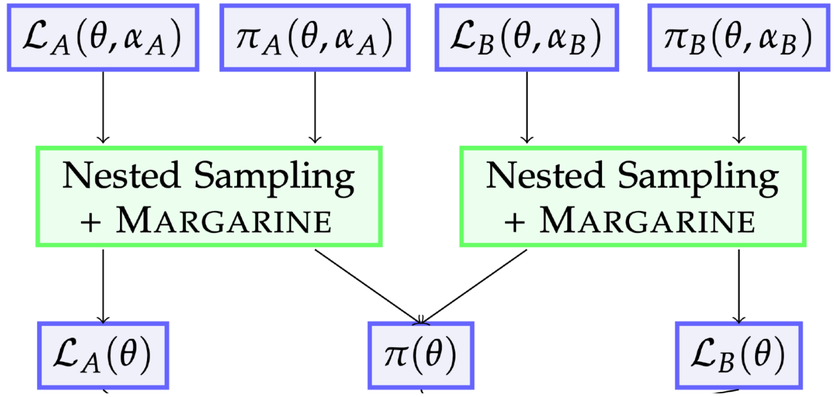
Joint Analysis and Hierarchical Modelling

Efficiently joint analysis

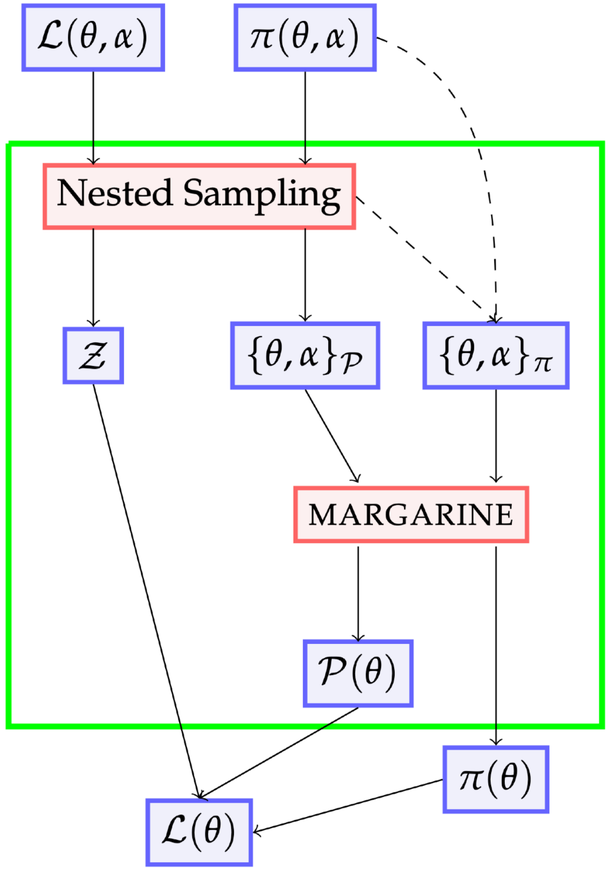


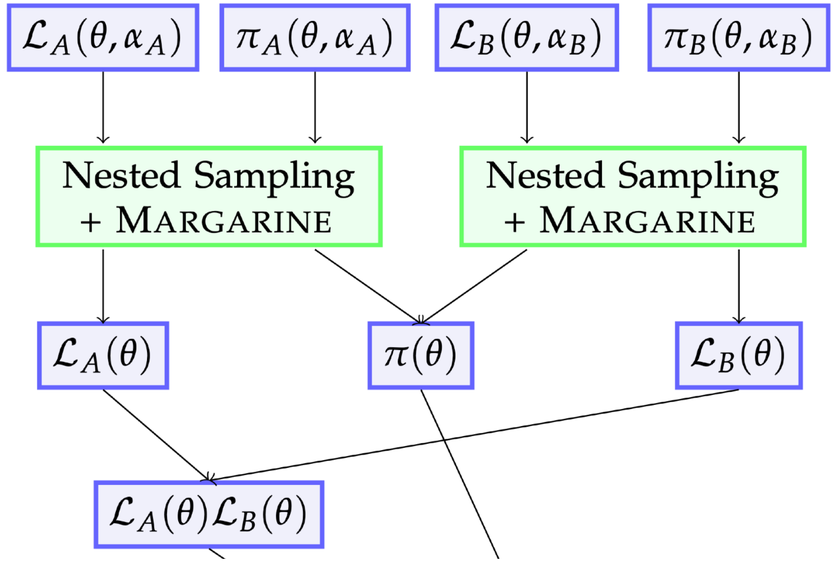
Efficiently joint analysis



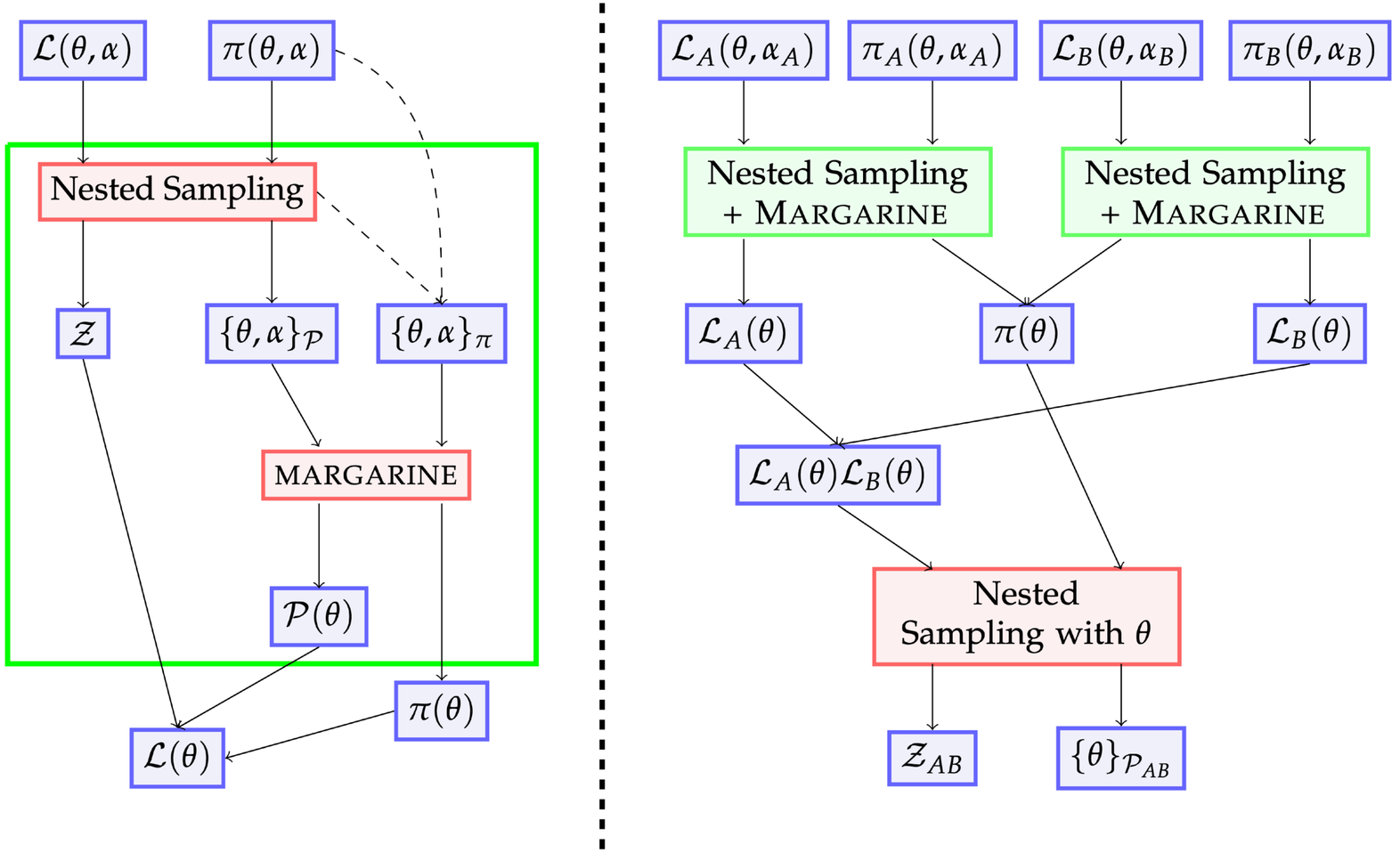


Efficiently joint analysis



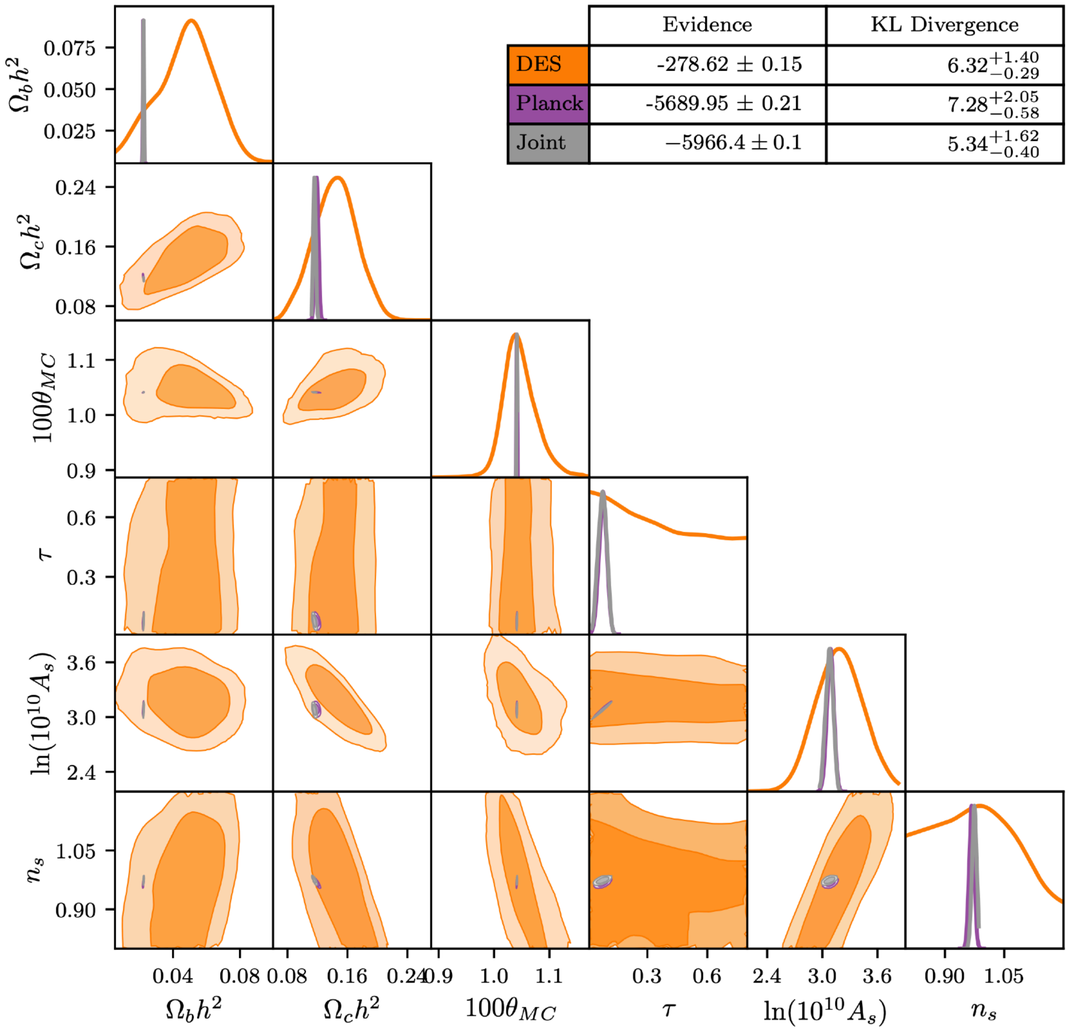


Efficiently joint analysis



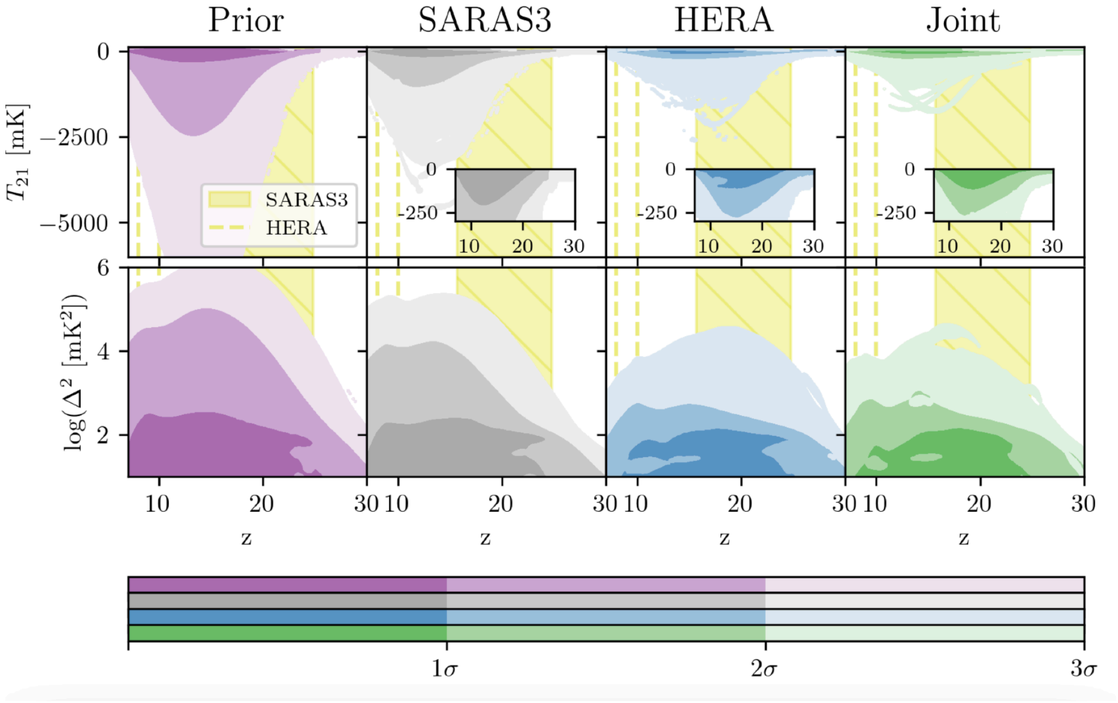
Example 1: Efficiently combining Planck and DES

- So using normalising flows we can emulate $P(\theta)$, $\pi(\theta)$ and $L(\theta)$ for a given θ
- $L_{Pla.+DES}(\theta) = L_{Pla.}(\theta)L_{DES}(\theta)$
- Reduce dimensionality significantly
- For Nested Sampling the runtime scales as $T \propto d^3$
- From $T \propto 20^3 + 26^3 + 46^3$ to $T \propto 20^3 + 26^3 + 6^3$
- Correct evidence and no double counting of priors



Example 2: Efficiently combining SARAS and HERA

- We also applied this to combine for the first time constraints from the sky-averaged 21cm signal and the associated power spectrum
- Showed that we could emulate the marginal likelihoods from SARAS3 and HERA and get some of the tightest constraints to date on the magnitude of the signal
- Work continued by Simon Pochinda [2312.08095] and Thomas Gessy-Jones [2312.08828]

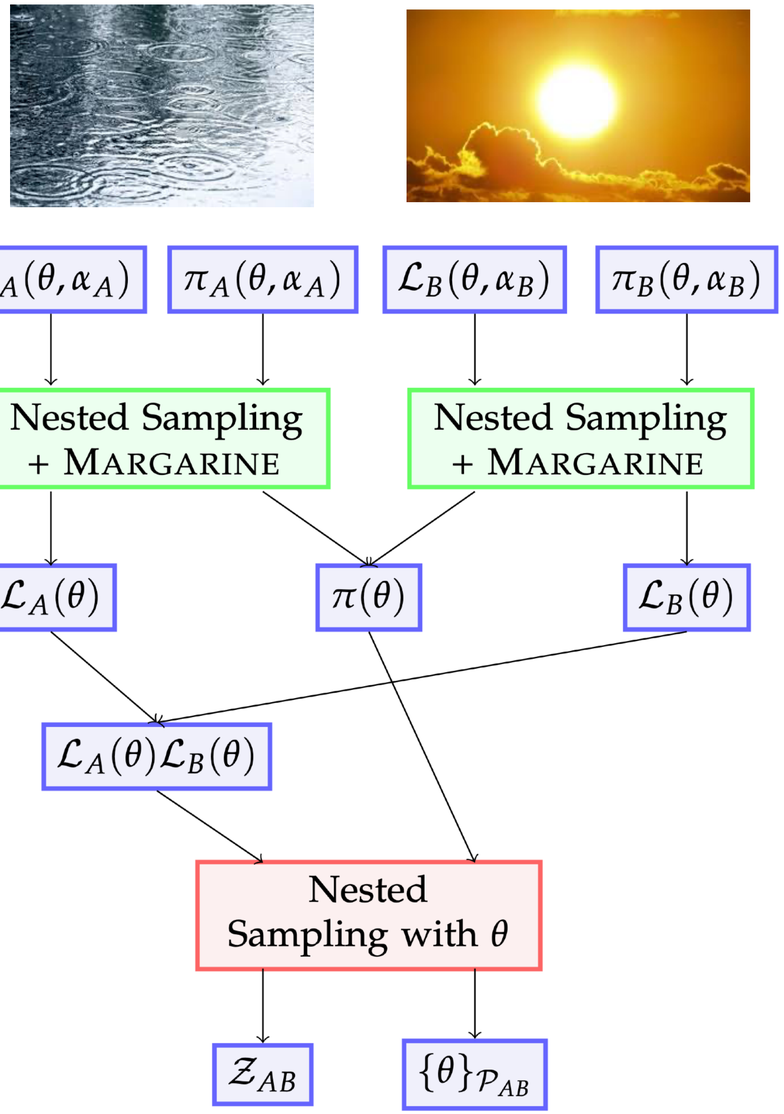


Example 3: Time dependent effects in 21cm

- Dipole antennas used for 21cm cosmology look down with as much power as they look up
- And the soil acts as a dielectric
- We can mitigate this a bit with a ground plane
- However, we still need to carefully model the properties of the soil
- We do this by modifying the model of antenna directivity

Example 3: Time dependent effects in 21cm

- Work led by Joe Pattison [2408.06012]
- Showed that we can combine observations from different time bins with different soil properties using ***margarine***
- Basic idea is to fit each time bin with a different beam based on our understanding of the soil at that time
- Then combine the constraints from the different fits

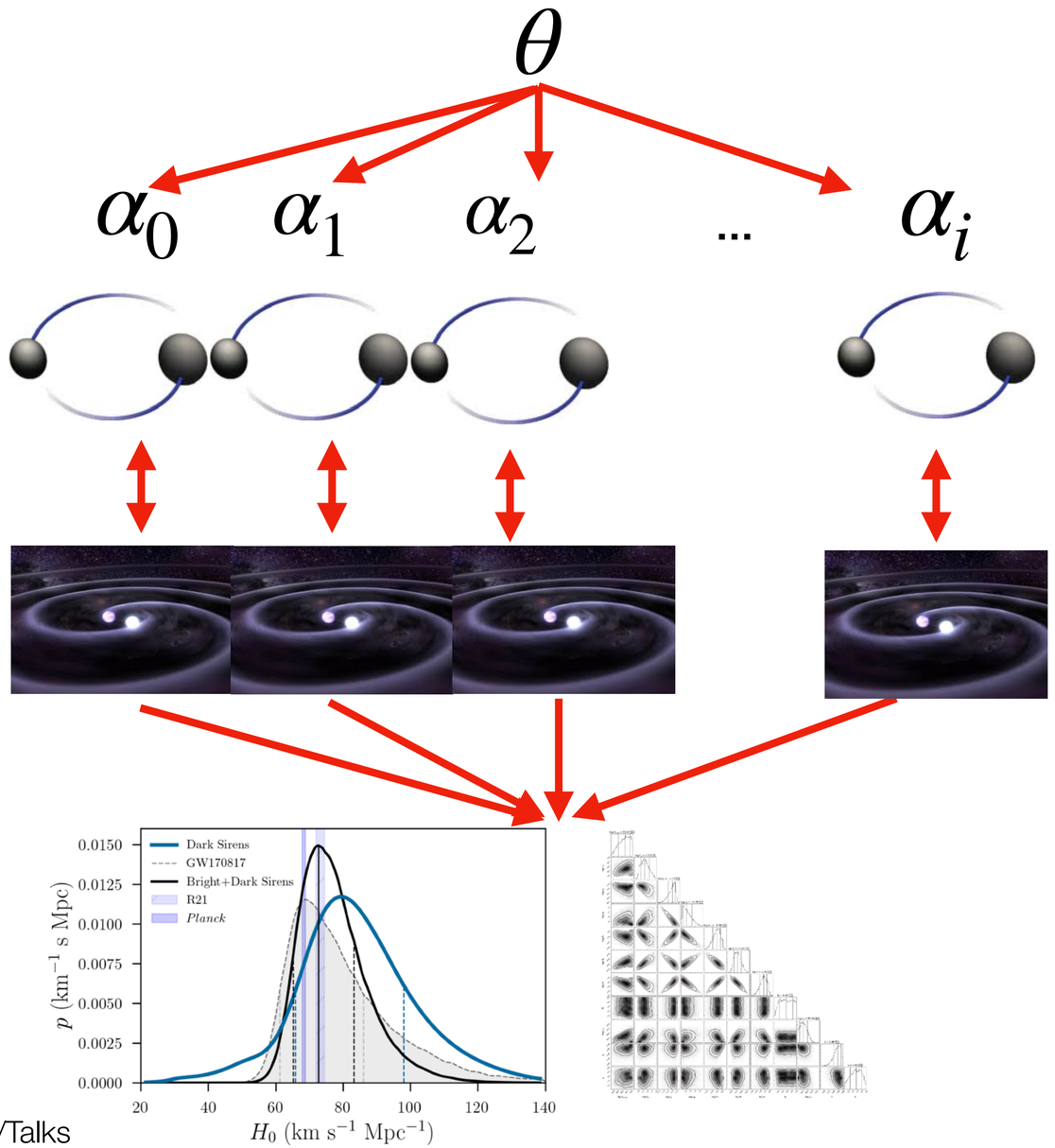


Example 4: Marginal likelihood for population level inference

- Hierarchical modelling we have population level parameters θ and individual object parameters α
- Propose fitting each object (usually do this anyway) to get $P_i(\theta, \alpha \,|\, D_i, M)$
- Using margarine to evaluate $P_i(\theta \,|\, D_i, M) \rightarrow L_i(\theta)$
- Then assuming the objects are independent sample the likelihood

$$L(\theta) = \prod_i^{N_{gal}} L_i(\theta)$$

- For traditional approach if $N_\theta = 6$ and $N_\alpha = 5$ then for 10 waves $N_{dim} = 56$ and $T \propto 56^3$

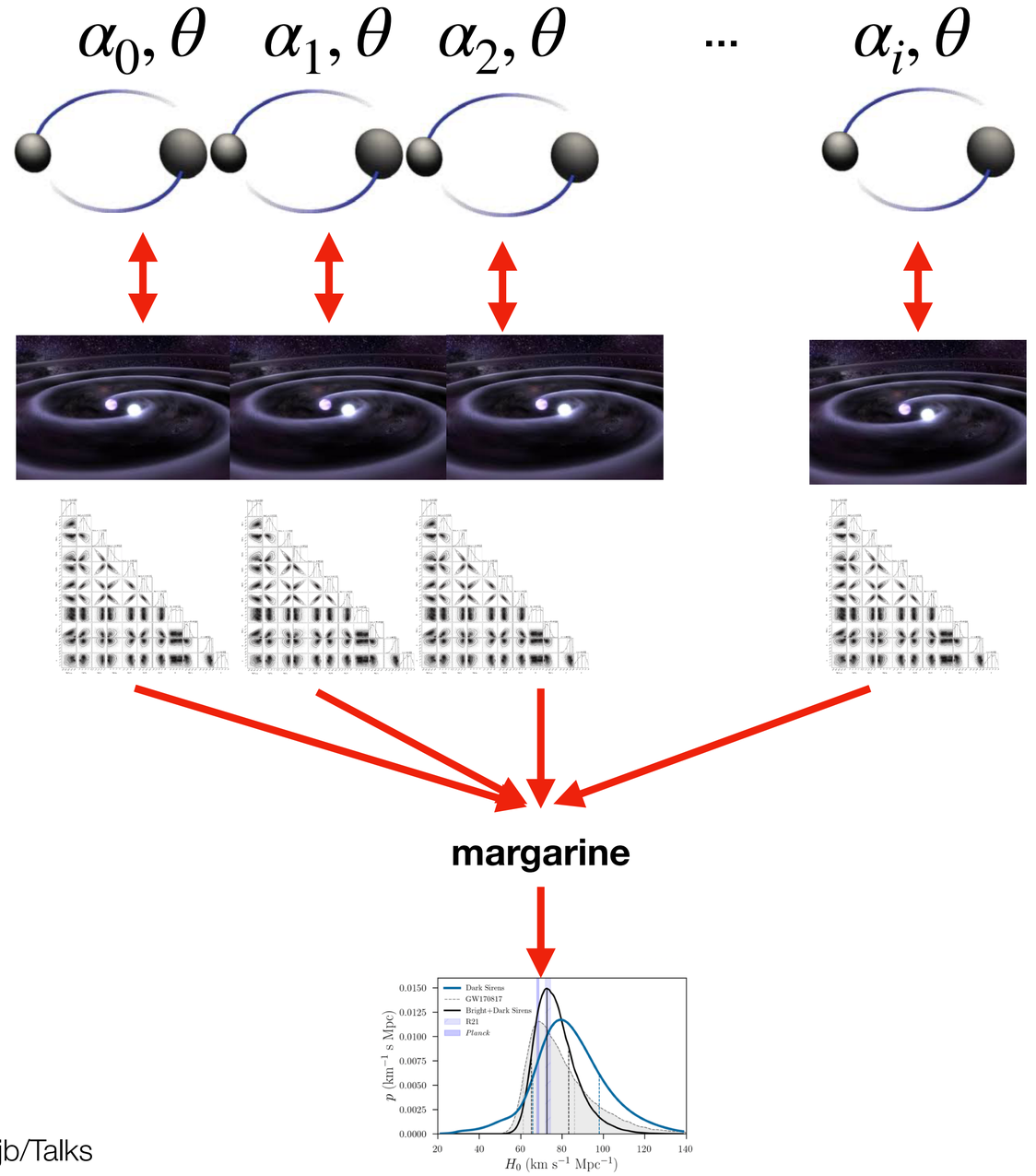


Example4: Marginal likelihood for population level inference

- Think about gravitational wave constraints on H_0 from dark and bright sirens
- For each object we can derive $P_i(H_0, M_1, M_2, S_1, S_2, \ldots | D_i, M)$
- We can then train flows on $P_i(H_0 | D_i, M)$ and $\pi(H_0)$ to evaluate $L_i(H_0)$
- And combine many observations to get

$$L(H_0) = \prod_i^{N_{GW}} L_i(H_0) \rightarrow P(H_0)$$

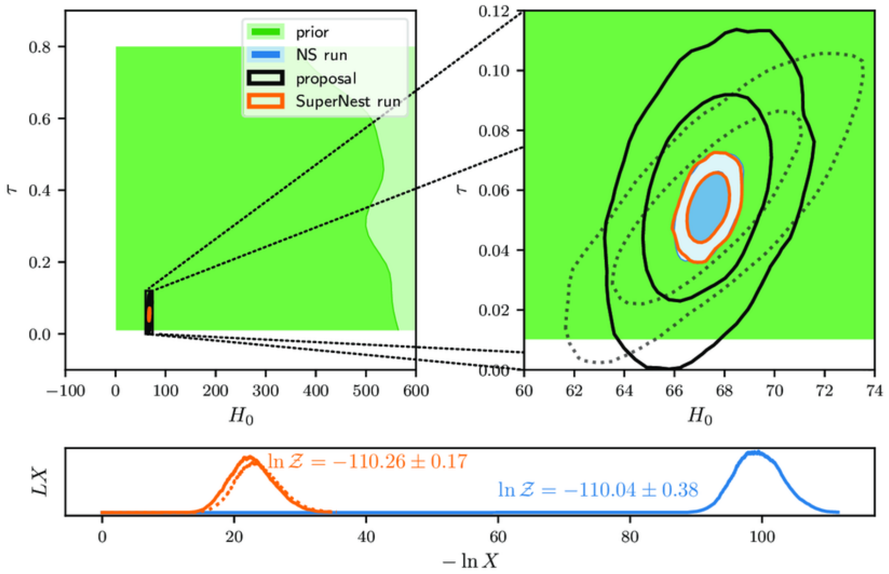
- $T \propto 10 \times 11^3 + 6^3 < 56^3$



Conclusions

What I did not cover

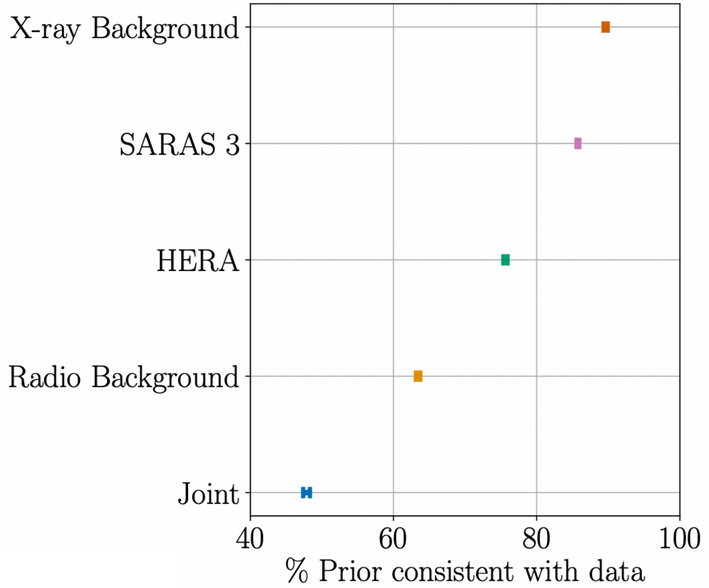
- Posterior repartitioning with normalizing flows and temperature dependent normalizing flows
- Mutual information with neural ratio estimation
- Population level analysis of galaxy surveys with SBI
- Radio cosmology and the REACH experiment!



[2210.07409]

What I did talk about

- Normalising flows (more generally density estimation tools) are incredibly useful for
  - interpreting our data
  - getting the most out of our inference products
- We wrote an extensible package called ***margarine*** that features a lot of the functionality I have been talking about
- Lots of papers including 2207.11457, 2205.12841, 2408.06012 and 2301.03298



margarine: Posterior Sampling and Marginal Bayesian Statistics

Introduction

margarine:	Marginal Bayesian Statistics
Authors:	Harry T.J. Bevins
Version:	1.2.8
Homepage:	https://github.com/htjb/margarine
Documentation:	https://margarine.readthedocs.io/

docs passing launch binder astro.IM arXiv:2205.12841