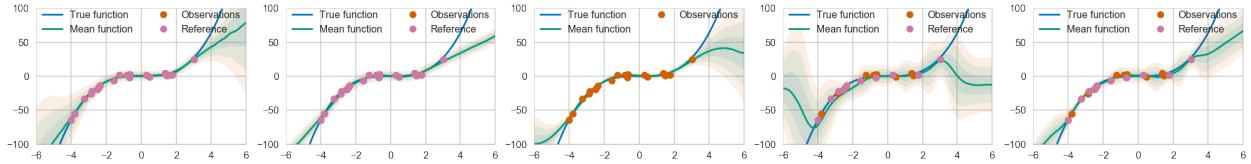
We would like to thank the reviewers for spending some of their time in thoroughly reviewing our work. Both R1 & R2 consider the work to be novel and significant, a fact that identifies the modelling framework of FNPs as a useful and important contribution to the community. The main negative points were in the experimental evaluation side and we will use this rebuttal as a way to address most of them.

The predictive distribution at eq. 12 is correct and a byproduct of the posterior approximations that we employed. If we instead employ the true posterior distribution over the latent variables, the predictive depends on the entire training dataset as $\mathbf{u}_R$ directly affects all of the points in $\mathcal{D}$ (see eq. 16 in the appendix). For this work we aimed for very simple posterior approximations and we left more involved ones as a point for future research. We further provide in the figure below an alternative toy regression fit with the same architectural details where we can see similar trends to the one provided in the main paper. All of the models were trained till convergence.



(a) MC-dropout (b) Neural Process (c) Gaussian Process (d) FNP (e) FNP$^+$

For all of the experiments in the paper, the NP was trained in a way that mimics the FNP, albeit we used a different set R at every training iteration in order to conform to the standard NP training regime. More specifically, a random amount from 3 to $num(R)$ points were selected as a context from each batch, with $num(R)$ being the maximum amount of points allocated for R. For the toy regression task we set $num(R) = N - 1$.

After the suggestions from the reviewers we also performed several additional experiments; we trained an NP with the same fixed reference set R as the FNPs throughout training, a standard variational Bayesian neural network with Gaussian priors / approx. posteriors over the weights and an FNP+ where we randomly sample a new R for every batch (akin to the NP and) and use the same R as the NP for evaluation. The results from these models can be seen at the table below and, as we can see, the FNPs still provide robust uncertainty while the randomness in R usually improves the o.o.d. detection, possibly due to the implicit regularization. We also included the results (mean & standard error) from both FNPs obtained after 5 replications with different Rs of $num(R) = 300$ (the one used in the paper); the error bars are larger on CIFAR 10 than MNIST, possibly due to it being a harder task. It should also be mentioned that the primary motivation for FNPs was not meta-learning, although it is something we plan to explore for future work.

	VI BNN	NP fix$R$	FNP+ rand$R$	FNP 5repl.	FNP+ 5repl.
MNIST	0.02 / 0.6	0.01 / 0.6	0.02 / 0.8	0.02± 0.0 / 0.7±0.0	0.02±0.0 / 0.7±0.0
nMNIST	1.33 / 99.80	1.09 / 99.78	2.20 / 100.0	1.95±0.06 / 99.93±0.03	1.97±0.05 / 99.97± 0.02
fMNIST	0.92 / 98.61	0.64 / 98.34	1.58 / 99.78	1.69±0.05 / 99.43±0.10	1.63±0.04 / 99.58±0.07
Omniglot	1.61 / 99.91	0.79 / 99.53	2.06 / 99.99	1.88±0.04 / 99.86±0.04	1.85±0.06 / 99.90±0.03
Gaussian	1.77 / 100.0	1.79 / 99.96	2.28 / 100.0	1.95±0.14 / 99.81±0.16	2.07±0.02 / 99.98±0.02
Uniform	1.41 / 99.87	1.42 / 99.93	2.23 / 100.0	1.99±0.06 / 99.96±0.02	1.95±0.06 / 99.96±0.02
CIFAR10	0.06 / 6.4	0.07 / 7.5	0.09 / 6.9	0.17±0.01 / 7.5±0.08	0.08±0.01 / 7.3±0.04
SVHN	0.45 / 91.8	0.46 / 91.5	0.56 / 91.4	0.86±0.05 / 90.74±0.81	0.51±0.04 / 91.3±0.76
tImag32	0.52 / 91.9	0.55 / 91.5	0.77 / 93.4	1.22±0.02 / 94.49±0.29	0.69±0.02 / 92.6±0.39
iSUN	0.57 / 93.2	0.60 / 92.6	0.83 / 94.0	1.33±0.02 / 95.71±0.24	0.75±0.02 / 93.8±0.38
Gaussian	0.76 / 96.9	0.20 / 87.2	1.23 / 99.1	1.05±0.10 / 93.73±1.29	0.60±0.08 / 93.6±1.09
Uniform	0.65 / 96.1	0.53 / 94.3	0.90 / 97.2	0.85±0.16 / 89.43±4.20	0.61±0.11 / 93.4±1.89

As for the comparison to variational / deep GPs; unfortunately, due to lack of time we cannot perform experiments with these baselines. On a theoretical level, the FNPs can be more scalable due to not having to invert a matrix for prediction. Furthermore, they can easily support arbitrary likelihood / noise models (e.g. for discrete data) in a straightforward way, in contrast to GPs where we have to consider appropriate transformations / warpings of a Gaussian distribution that usually require further approximations.

Similarities with ANP; indeed we did mention it in the related work section. We can view the attention weights of ANP as providing the graph edge probabilities in FNP. The key difference between the FNP and NP still applies for the ANP though, as ANP still has a global latent variable. Unfortunately, due to lack of time we cannot perform experiments with ANP, but we believe that the additional results we presented in this rebuttal make FNP contribution convincing.

Mapping for $\mathbf{u}$; indeed the type of this mapping is important and the reason why appropriate regularization is necessary. This is also one of the reasons why we tied most of the model / variational parameters, as mentioned in section 2.2. It should be mentioned that NPs can have similar drawbacks, e.g. how do you select the appropriate way to incorporate the global context θ in the prediction. Having said that, we can also view it as a capability of the FNPs, since it can serve as a knob that can be adjusted for the specifics of a given application.