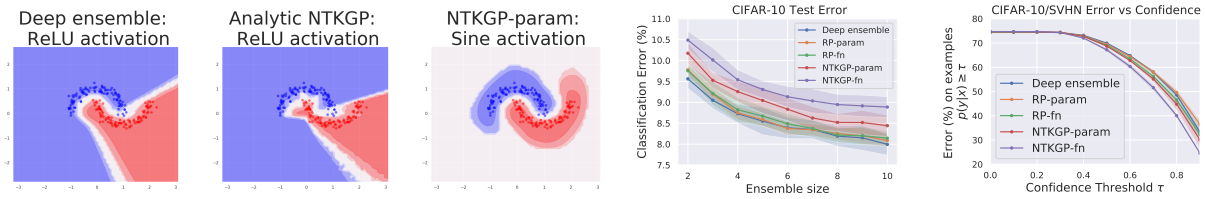
We thank reviewers **R1**, **R2**, **R3**, **R4** for their time and constructive reviews on our submission, which we will incorporate to improve our paper. Due to limited space, we will only be able to address the major points from the reviews:

Benefits of a GP posterior ensemble interpretation (addressed to **R1** & **R4**) We agree with **R2** that a key strength of our work is that it "provides a formal treatment of the relationship between deep ensembles and Bayesian posterior predictive distributions". Posterior inference offers a principled way to convert prior beliefs into predictive uncertainties, and provides o.o.d. robustness via Bayesian marginalisation [14]. Moreover, Bayesian ML has a rich history [3] and is an active research frontier. One practical benefit to the GP posterior interpretation is selecting hyperparameters, like activation, of the NTK/NN architecture (akin to choosing GP kernel) that best model prior beliefs about data. For example, the NNGP/NTK correspondence allows one to deduce that Sine activation can alleviate overconfidence (see [43]) of ReLU deep ensembles on Two Moons classification. This is because the ReLU kernels do not decay away from the training data, as can be seen in Eqns S14,15 of [21], unlike the Sine kernels, as can be seen in `stax.py` of the Neural Tangents library [31]. In the left 3 plots below, we demonstrate this empirically (with two layer NNs of width 500, MSE trained with scaled one-hot regression targets and no observation noise, which are then fed into cross entropy to get probabilities): we see that both deep ensembles [11] and NTKGP analytic (with small noise $\sigma^2>0$ added for numerical stability, for **R4**) are overconfident with ReLU activation (denoted by blue and red shaded regions), but NTKGP-param with Sine activation has low confidence (white regions) away from the training data (points), as desired.



Additional experiment on CIFAR-10 (**R2**, **R3**) We repeated the MNIST/NotMNIST experimental setup using the Myrtle-10 CNN with 100 channel width (Shankar et al., arXiv:2003.02237) trained on CIFAR-10 with SVHN o.o.d. test set. We changed our classification methodology to use MSE loss (using scaled one-hot regression targets with scale selected via moment-matching with NTK, small $\sigma^2>0$) before temperature scaling on a validation set. In the right 2 plots above, we see that our NTKGP methods perform slightly worse on in-distribution test accuracy (<1% higher error), but outperform all baselines on o.o.d. detection in the Error vs Confidence plot (far right). For instance, NTKGP-fn exceeds baselines by between 8-10% accuracy on (combined in-dist+o.o.d.) test points with confident predictions (e.g. confidence τ=0.8). This o.o.d. performance gain is crucial for safety-critical applications (e.g. self-driving cars).

Computational overhead (**R1**, **R2**, **R4**) We would like to clarify that for a training set of fixed-size (e.g. CIFAR-10) the train-time overhead of our methods is negligible compared to standard deep ensembles [11]: one can obtain and store our fixed additive JVPs δ in a single pass over the training data. This was mentioned on lines 528-531. For test-time constrained applications, we could apply distillation techniques, as is common with standard deep ensembles.

Novelty (**R1**, **R4**) Though presentation has been simplified, we respectfully disagree that the novelty of our paper is straightforward. We are (to our knowledge) the first to consider GP(0,NTK) prior instead of GP(0,NNGP), in order to align posterior inference with optimisation of *all* NN layers. Our contributions are distinct to, not extensions of, [22, 23], and give different limiting predictive distributions to [22], see Table 1. Also, if there is no modelling of observation noise, σ^2=0, then RP-param [22, Eq. 4] and anchored ensembles with MSE [23, Eq. 8] become standard deep ensembles [11]. On the other hand, NTKGP-param still retains its posterior interpretation (Corollary 1), using fixed additive JVP corrections with no regularisation nor noisy targets. This is the case in the two moons ensembles above. It is only when modelling observation noise that we synergise our methods with [22].

Individual responses (**R1**) We contest the "marginal" empirical improvement of our work: Figure 3 (right) depicts significant gains of our methods for o.o.d. NotMNIST detection over baselines. The NTK & standard parameterisations are introduced in Appendix A. (**R2**) We believe the slightly worse in-distribution test performance of our methods can be alleviated with thorough NTK hyperparameter tuning. (**R3**) When modelling observation noise, $\sigma^2>0$, our regularisation scheme (Appendix D) enables closer alignment to the kernel regime in standard parameterisation (lines 492-499) and nullifies problems caused by the fast decay of NTK eigenvalues (lines 84-86). (**R4**) $\Theta \succeq \mathcal{K}$ follows from the NTK being a sum of p.d. contributions from different layers and $\mathcal{K}$ is the contribution from last layer, see Eq. S29 of [21]. JVPs are more memory-efficient in forward-mode than reverse-mode AD; we will add this. Please see lines 511-513 for discussion of parameter and function space methods; our code is open-source and we are working with the Neural Tangents authors [31] to integrate our work. It is unclear if analytic NTKGPs are preferable to analytic NNGPs when both are tuned, due to cost and predictive-mean performance (see §3.2 of Lee et al., arXiv:2007.15801); we focus on giving a posterior interpretation to deep ensembles for wide but finite NNs, and lack the compute needed for comparisons of large-scale analytic NTKGP/NNGPs. For the prediction decomposition, setting $\Theta^{\sigma}_{\mathcal{X}\mathcal{X}} = \Theta(\mathcal{X},\mathcal{X}) + \sigma^2 I$, we obtain:

$$\tilde{f}_\infty(\boldsymbol{x}^*) = \Theta(\boldsymbol{x}^*,\mathcal{X})\left[\Theta^{\sigma}_{\mathcal{X}\mathcal{X}}\right]^{-1}\mathcal{Y} + \underbrace{f_0(\boldsymbol{x}^*) - \Theta(\boldsymbol{x}^*,\mathcal{X})\left[\Theta^{\sigma}_{\mathcal{X}\mathcal{X}}\right]^{-1} f_0(\mathcal{X})}_{f_0} + \underbrace{\delta(\boldsymbol{x}^*) - \Theta(\boldsymbol{x}^*,\mathcal{X})\left[\Theta^{\sigma}_{\mathcal{X}\mathcal{X}}\right]^{-1}\delta(\mathcal{X})}_{\delta}$$