We thank all reviewers for their careful reading and their detailed and constructive comments. We appreciate the positive feedback of all reviewers testifying our approach to be "advancing the field of deep Gaussian processes" (R2) and to inspire future research in this area (R1,R2). In particular, we want to thank the reviewers for acknowledging the constructive and detailed proof (R2), the easy to follow, well-written and well-contextualised manuscript (R1-R4), the careful consideration of computational aspects (R1,R3), and the helpful, detailed appendix (R3).

We first address the shared reviewer comments and then individual ones. The paper will be revised accordingly taking also further minor comments and suggestions of the reviewers into account.

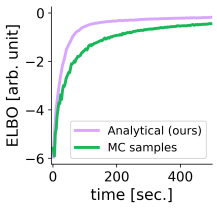
Empirical evaluation (R1-R4) We agree with the reviewers that the presentation of the results was not entirely convincing. This is mainly due to the random 1D-projection of the extrapolation experiment: The direction of the projection has a large impact on the difficulty of the prediction task. Since this direction changes over the repetitions, the corresponding test log-likelihoods vary considerably, leading to large standard errors that hampered the comparison between the methods. We resolved this by performing a direct comparison between MF and STAR DGP as proposed by R1: To do so, we computed the frequency of test samples for which STAR DGP obtained a larger log-likelihood than MF DGP on each train-test split independently. Average frequency μ and its standard error σ were subsequently computed over 10 repetitions. On 5/8 datasets STAR DGP significantly outperforms MF DGP ($\mu > 0.50 + \sigma$), while the opposite only occurred on *kin8nm*. As suggested by R2, we also compared MF to FC DGP leading to similar results (see new table).

Dataset	MF vs. STAR	MF vs. FC
boston	**0.55 ± 0.04**	**0.58 ± 0.04**
energy	**0.73 ± 0.05**	**0.70 ± 0.05**
concrete	**0.57 ± 0.04**	**0.56 ± 0.02**
wine red	**0.57 ± 0.04**	**0.57 ± 0.03**
kin8nm	*0.36± 0.03*	**0.59 ± 0.02**
power	0.44 ± 0.06	**0.68 ± 0.03**
naval	**0.67 ± 0.06**	*0.24 ± 0.07*
protein	0.49 ± 0.03	0.50 ± 0.01

Intuition for structured approximation (R1) When we started working on the topic, we had the hypothesis that structured approximations would be especially helpful for test points that are distant from the training data and this idea also guided the layout of our experiments. While the results in our new table and Fig. 2 support our hypothesis, we were neither theoretically nor empirically able to pinpoint the underlying mechanism. We agree with R1 that an examination of inner layer samples for different structures (similarly as done in Ref. [34]) and the corresponding effects on the outputs are important research questions that need to be addressed in the future.

Train-test split (R2) We are the first to study the extrapolation behaviour of DGPs. While we agree that the splitting criterion could be improved, our experiments already reveal that established DGPs struggle in this setting. Furthermore, we indeed used the standard conventions for creating Tab. S2 and will move it to the main paper to facilitate comparison to related work.

Convergence analysis (R2) We thank the reviewer for proposing an empirical comparison of the convergence speed between analytical and MC marginalization. As proposed, we maximised the ELBO with both algorithms (using FC DGP L3 on the *concrete* dataset). We confirmed that the analytical marginalization converges quicker in terms of runtime (see new figure).



Choice of structural approximation (R2) In addition to the empirical motivation of our STAR structure (Fig. 1), the stripes pattern can also be justified from the model architecture: We expect the residual connections, realised by the mean functions (footnote 2), to lead to a coupling between successive latent GPs. In general, choosing the optimal structured approximation is highly model and data dependent. We agree that this is an important aspect of future work.

Intuition for proof (R3) We are amazed to find this heuristic argument in our reviews. While mathematically not rigorous, it provides the correct intuition. In fact, it was precisely the same reasoning that initially allowed us to come up with the induction hypothesis (Lem. 2). We will include this argument in the final version to provide additional guidance.

Code and experiments (R3) We thank the reviewer for the positive feedback on our unit tests. As suggested, we will also make the source code for the experiments publicly available. Test log-likelihoods were computed on the marginals.

Inconsistency between coupled posterior and factorised prior (R4) We agree that the role of coupled priors has not been thoroughly studied in deep GPs and should be investigated in more detail as it is done for the weight prior in Bayesian neural networks [e.g. Wenzel et al., ICML 2020].

Relationship between variance and MSE (R4) For a calibrated method, the predictive variance σ_i^2 is the expectation of the squared error (SE_i) for test sample i. We estimated the latter by the empirical mean squared error (MSE) of test samples with a similar σ_i^2. The predictive variance σ_i^2 and the empirical SE_i are also compared in the test log-likelihood, $\log \mathcal{L} = -\frac{1}{2} \sum_i \left(\log \sigma_i^2 + \frac{\mathrm{SE}_i}{\sigma_i^2} \right)$, in which inaccurate predictions are penalised by the first and overconfident predictions by the second term (cf. the quantities in Fig. 2).

Number of layers (R4) We agree that the improvement of adding more layers (L2 to L3) in Tab. S2 is only significant for the *protein* dataset. However, this is in line with the results published in [24, Tab. 7], where the largest improvement is also observed on *protein*, and the only other dataset with significant but considerably smaller improvement is *kin8nm*.