We are grateful for the reviewers' detailed comments. We will rectify all small writing errors, and offer comments for more significant issues below.

@ R1 – **Other Gaussian baselines:** We concentrated on full-rank Gaussians because these give dramatically better results than diagonal Gaussians; we will make sure to clarify this earlier on. With minimal space, we did not want emphasize the issue of full-rank vs. diagonal Gaussians, since this is a (relatively) well covered issue. In retrospect, however, we definitely appreciate your point that this would be valuable to see in the context of our full experiments. We will at a minimum perform a comparison to diagonal Gaussians and add the results to the appendix. (We are concerned about a lack of space for a full discussion in the main paper.) The summary is that diagonal Gaussians perform much worse!

@ R1 – **Time tradeoffs :** Thank you for pointing out that this discussion needs attention. Both of the your suspicions are correct. First, in preliminary experiments, we found an increase in the number of coupling layers correlated with better performance. So there is indeed a time-quality tradeoff there. (We do not attempt to find the best tradeoff as our goal is not really to find the best possible flow.) This should be clarified. Second, it is true that the Gaussian scales as $O(n^2)$ while flows scale as $O(n)$. After some calculation, one can show that the number of parameters for Gaussian will exceed that for RealNVP if the latent dimensions exceed 2000; (we discuss this briefly in Section F, the last paragraph.) While flows scale better, the computational constant is also higher. At the same time, when the dataset is large, the cost of $\log p(z, x)$ dominates costs involving the variational distribution. And further, our code is implemented in Python while models are implemented in Stan, which complicates the meaning of running times in low dimensions when interpretation overhead is a factor. Still, we accept that running times measurements are helpful even if provisional, and will add these to the appendix.

@ R1 – **ELBO correlation with other metrics:** Of course, an ELBO improvement is equivalent to an improvement in KL divergence between a variational approximation and the true posterior. Several works have reported an empirical correlation between the ELBO improvements with improvements in test-likelihoods (see [21, Appendix, Table 3, 4, and 5]; Tao et al., 2018, Figure 4; Mishkin et al., 2018, Figure 4.) and accuracy of posterior moments (see [9, Figure 6]; [10, Figure 7, and Figure 9 to 22]).) We expect our methods to follow similar correlations for ELBO improvements. The advantage of using an ELBO is that the results are far more stable since they do not depend on details like train-test splits, or the particular metric of accuracy.

@ R1 – **IAF-style flows and Polyak-Ruppert averaging:** Yes, as suspected, we use RealNVP because it is reasonably "generic" and supports the STL estimator; we will make this clear in the paper. We agree that a more exhaustive comparison of possible flows would indeed be useful, but may not be the highest priority given space limitations. We are intrigued by the hypothesis about Polyak-Ruppert averaging, but must be honest that it may be challenging to include.

@ R2 – **Assessment of Variability:** Thank you for mentioning the challenge of statistical measure of variability in empirical CDFs. We considered this issue at length while writing the paper. There are standard methods for calculating confidence intervals, but applied naively, and these would not do what we want. (They would estimate generalization to other "models" (drawn from the same distribution of models) rather than generalizing to running the same inference methods with different random numbers.) Instead, we settled on the simplicity and transparency of running three completely independent experiments for each change and superimposing the results. We believe this is adequate because the variability is minimal in almost all cases. (The only exceptions are due to the ADVI step size scheme; however, the improvements are so vast that the conclusions are not in doubt.) We do provide full results for all independent trials in Appendix J and will add more discussion to the paper.

@ R3 – **Prior Work and writing suggestions:** Thank you for pointing out the two missed papers–we will discuss them. We will also make the Broader Impact section more specific. Regarding the "default choice," it is certainly true that a small number of models exist on which the changes degrade performance. We will better acknowledge this. However, we suspect that it may be difficult to draw reliable conclusions about what models are at issue. Concretely, compare ADVI against the best-performing approach (4c). The set of models where ADVI performs better is precisely one: `gp-predict`. We certainly agree that our analysis is not the "last" word. We think that with more innovation in analysis techniques (and a larger corpus) model-family specific details can be teased-out; still, this paper represents a first step in the direction of rigorous empirical evaluation for inference research. Lastly, we will make the discussion in Appendix C more prominent.

Tao et al. Variational Inference and Model Selection with Generalized Evidence Bounds. ICML, 2018

Mishkin et al. SLANG: Fast Structured Covariance Approximations for Bayesian Deep Learning with Natural Gradient. NeurIPS 2018.