We thank the reviewers for their useful and thoughtful feedback. We are glad to see that our work was found "*highly relevant to the NeurIPS community*" (**R1**) and to be addressing "*a well-motivated problem*" (**R3**) in "*an important area*" (**R1**); that our empirical validation was "*sound*" (**R1**), "*comprehensive*" (**R2**), "*extensive: 7 models are considered and two state-of-the-art baselines are included*" (**R3**), and "*thorough* (**R4**)"; with our claims "*supported by a comparatively large set of real-world experiments*" (**R1**), results that "*largely align with the claims*" (**R2**) and "*are significant to support the claims of the paper*" (**R4**). That is, "*the experiments show that the proposed method outperforms the baselines and works consistently well in all the real-world models considered*" (**R3**). We address the reviewers' comments below and incorporated all feedback in the revised paper. Here with 'VBMC' we refer to our method for noisy inference.

Known noise (R1, R4). The assumption that log-likelihood noise σ_obs is (approximately) known is less limiting than it might seem. The common *synthetic likelihood* approach already needs multiple samples to build the multivariate normal pdf used to estimate the likelihood, hence σ_obs can be easily estimated via bootstrap with negligible cost (as shown in [10]). *Inverse binomial sampling* (IBS, [41]), used extensively in our analyses, automatically provides both an estimator of the log-likelihood and of σ_obs for free. For other techniques, bootstrap is often a feasible and easy-to-implement choice. **R4:** "*How does the estimation error* [of σ_obs] *affect the results? Will it change the claims of the paper?*" Great question. We performed a new analysis by rerunning VBMC on several problems (aDDM, Timing, Neuronal) while drawing the estimated $\widehat{\sigma}_\text{obs} \sim \text{Lognormal}\left(\ln \sigma_\text{obs}, \sigma_\sigma^2\right)$ for increasing values of noise-of-estimating-noise, $\sigma_\sigma \geq 0$. At worst, the performance of VBMC via the MMTV metric degrades only by ~ 0.03 points on average (e.g., from 0.13 to 0.16 on the Timing problem; see Fig. 3 for reference), with σ_σ up to 0.4 (i.e., $\widehat{\sigma}_\text{obs}$ roughly between $0.5 - 2.2$ times the true value); showing that VBMC is very robust to imprecise estimates of the noise. Thus, our claims are unaffected.

Dimensions & Applicability (R1, R4). The limit of VBMC to $D \sim 10$ input dimensions is common to approaches that use GP surrogates (e.g., Bayesian optimization [22]), and an open area of research which we intend to pursue. Still, *plenty* of models in computational biology and neuroscience have up to ~ 10 parameters, so there is *wide applicability* of our method. As proof, consider the widespread usage within computational neuroscience and related fields of the recent BADS toolbox [26] for (noisy) Bayesian optimization, which shares similar limitations as VBMC in terms of D. Crucially, VBMC is a large step forward with respect to BADS for practitioners in the field in that it affords *full Bayesian inference* (as opposed to limited to point estimation), so we expect it to impact a wide audience. Finally, while our method does not directly apply to high-D machine learning models (e.g., Bayesian neural networks), it could be used to infer posteriors over *hyperparameters* of ML models, which we see as a relevant research direction.

Theoretical guarantees (R2, R4). While we share the reviewers' desire for convergence guarantees, we also note that convergence proofs of adaptive Bayesian quadrature methods are outstanding theoretical contributions in themselves. For example, despite a growing body of work on adaptive methods over the years, *only last year* Kanagawa and Hennig [37] were first able to prove convergence for a class of *local* acquisition functions in Bayesian quadrature. By contrast, our paper adds to the literature of solid empirical contributions with theoretically-motivated choices. We believe that our strong empirical validation, judged very positively by all reviewers, while not at all replacing a mathematical derivation, should provide confidence in our method and inspire future theoretical research on VBMC.

Motivation of a_VIQR acquisition function (R2, R3). The rationale for going from a_IMIQR (Eq. 7) to a_VIQR (Eq. 8) is: (1) sampling from the posterior $\exp(\overline{f}(\boldsymbol{\theta}))$ is relatively hard, whereas sampling from $q_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ (the variational posterior) is trivial; (2) $q_{\boldsymbol{\phi}}(\boldsymbol{\theta})$, by construction, approximates $\exp(\overline{f}(\boldsymbol{\theta}))$ up to a normalizing constant (irrelevant for optimization). Thus, we go from Eq. 7 to Eq. 8 by swapping $\exp(\overline{f}(\boldsymbol{\theta}))$ with $q_{\boldsymbol{\phi}}(\boldsymbol{\theta})$, with substantial gains (see Table 1). While a_VIQR adds a layer of variational approximation ($q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \approx \exp(\overline{f}(\boldsymbol{\theta}))$), hence the name, it still directly approximates a_IMIQR.

Response to remaining comments. **R1:** *Modelling* σ_obs, while possible, is tricky in practice due to the trade-off between σ_obs and variability of the latent function, which may be hard to disambiguate. However, as argued above, obtaining (approximate) *estimates* of σ_obs is very often a viable solution; we leave modeling of *unknown* σ_obs as future work. **R2:** In terms of novelty, we propose *variational whitening* (which improves performance on hard problems); we extend (in a fairly straightforward way) previous work from [10,11] to the VBMC framework; we build a novel, extensive noisy benchmark with many models and real datasets from computational and cognitive neuroscience, showing that VBMC vastly outperforms state-of-the-art, providing a meaningful contribution to the field. **R3:** We discussed potential reasons of failure for WSABI and GP-IMIQR in the Supplement, D.2. We believe we performed a fair comparison, in that for both WSABI and GP-IMIQR we tried to fix potential issues, and report the best-performing variants. Re. multimodal posteriors, VBMC (similarly to MCMC) would have trouble with far, disconnected modes, and it is an active research area; the exploratory tendency of GP-IMIQR might help here, but also lead to instabilities. Re. the noiseless case, see Fig. S6, in which we varied the amount of noise; interestingly, most acquisition functions perform similarly with no noise ($\sigma_\text{obs} \approx 0$). In Eq. 7, the prior density is included in $\exp(\overline{f})$ (f models the log-joint). **R4:** "*Lack of ablation study.* [...]" The ablation studies are mentioned in the text, and fully reported in the Supplement (E.3, 'Lesion study'). In particular, removing variational whitening degrades performance on the difficult Neuronal and Rodent problems. However, we found no major differences in performance on the other problems; leading us to claim that the improved performance of VBMC on noisy problems (wrt. the original framework [5]) is mostly due to the new acquisition functions.