We thank the reviewers for their insightful comments regarding our paper. R1 and R2 highlighted the technical quality and clarity of our work and the novelty of the application. All minor comments will be addressed in the revised paper. Here, we briefly reply to selected major points raised by the reviewers (references refer to the main paper):

Significance of the paper to the NeurIPS community (R3)
Our work shows that models for system identification, which are widely used in the NeurIPS community, can be supplemented by biophysically inspired components, such as a model of vesicular release at a ribbon synapse, and inference for such models can be performed in a Bayesian manner. Our additional analysis (see point 2 below) shows that our model clearly outperforms GLM-style models for the type of data studied here. On the technical side, we show that for such models which are fast in evaluating, already a simple ABC method can estimate the posterior distribution efficiently and no additional overhead such as training a DNN or GP - like in reference [6] and [8] - is necessary.

Comparison with previous models (R2)
To address this point, we performed additional analysis and compared the LNR-model to a GLM with stimulus and self-feedback term and Poisson noise [2]. In contrast to the LNR-model, the GLM was not able to capture the multiple vesicular release with more than three vesicles at a time and showed much larger discrepancies overall (18.3 ± 1.8, mean $\pm$ std compared to 6.5 ± 0.3 for the LNR-model). The weights of the linear part for the release history captured the suppression of additional release after a release event partly but could not model the full dynamics. This analysis demonstrates that taking biophysical constraints into account can dramatically improve prediction accuracy of system identification models. We will include an appropriate figure and details in the revised paper.

Parametrization of the model (R2, R3)
Stimulus filter:
The learned filter in the LNR-model is different from the filters recovered with e.g. the STA, as the release dynamics are taken into account in its estimation. For simplicity, we assumed a stimulus kernel with one parameter only, but a basis function approach [2] to allow for more flexibility could in principle be used as well. However, this would lead to a higher dimensional parameter space, making inference less efficient. Exploring this trade-off further is an interesting direction for future work.
Slope parameter:
The slope parameter k of the non-linearity is indeed underestimated, likely because of the "non-linear" effect of k on the slope of the non-linearity. Our method sets k to a value where a further increase would not change significantly the output of the model.
Summary statistics:
Indeed, the weights for the summary statistics were chosen to make some features more important, but our experiments suggest that inference results were largely insensitive to the exact choice. While we chose the weights heuristically, in principle, cross-validation could be used for a more systematic procedure.
We will improve our discussion of all three aspects in the revised version of the paper.

Form of the posterior and acceptance strategy (R3)
Initial experiments showed almost uncorrelated posterior distributions for most of the parameters. Hence we decided to factorize the distribution in most dimensions and modeled only the non-linearity parameters as a multivariate normal distribution. In general, the described formulas for the two dimensional multivariate distribution would indeed generalize straight forward to higher dimension. However, distributions such as the Γ-distribution for λ_c would then have to be approximated.
Using a fixed acceptance threshold for the loss results in inefficient updates of the proposal prior in early iterations as very few or even no samples are accepted in each round. Using an adaptive threshold might remedy this, but would likewise affect the estimated variance.

Runtime and Complexity (R1)
The runtime of the presented ABC method is dominated by the forward simulations of the model, with a complexity $\mathcal{O}(n)$ if n is the number of drawn samples. This complexity is similar to SNPE-B [6], which in addition requires training of a mixture density network, while we resort to analytic updating formula. Although for expensive simulations, either strategy is often only a small fraction of the total run time, our method should be advantageous if the simulation is fast and the posterior unimodal. As already mentioned in the main text, the direct estimation of the posterior stands in contrast to SNL [9] or BOLFI [8] where the inference of the posterior involves a second sampling step via MCMC which can be slow. In addition, BOLFI [8] uses a Gaussian process with complexity $\mathcal{O}(n^3)$ in the vanilla version to approximate the likelihood, which can be prohibitively slow. Additional discussion will be added to the paper.