We thank the reviewers for their positive and insightful feedback as well as the research ideas for future work (e.g. the LSTM experiments suggested by R4). 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):

Ablation study (R1 & R2) We further investigated the contribution of the different model parts by training two ablated models, one without AC feedback terms (a deterministic version of the LNR model in [24]) and one without release block but all AC feedback structure.

The ablated models all showed lower training correlation than the BCN model. In particular, they failed to capture certain features of the chirp response: For the LNR model, we found missing "feedback features" (like higher baseline in some On cells or missing responses to small amplitude On steps, Fig. 1A). For the BCN without release block, we found a mismatch in the response to the long On/Off phase of the stimulus as the adaptation processes could not be fully captured (Fig. 1B). Interestingly, the generalization performance of the ablated models was surprisingly good. In particular, the BCN model w/o release block showed high correlation on the natural movie dataset. We found that some of the functional properties of the release block can likely be captured by the feedback structure. Also, the natural movie data does not contain extended light steps, for which the special release properties of the ribbon synapse are so prominent (Fig. 1). We will add a discussion of these additional results to the paper.

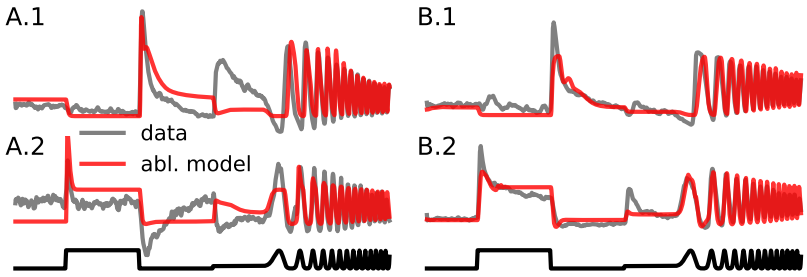

Figure 1: **Predictions for ablated models on the chirp stimulus. A.** BCN without feedback: prediction for global BC1 (A.1) and global BCR (A.2) response. **B.** BCN without release block: prediction for local BC4 (B.1) and local BC6 (B.2) response. Stimulus is shown in black.

Training vs model structure (R1 & R4) Comparing the influence of training and model structure is non trivial. As a first step, we ran the best BCN model, kept the stimulus filter and release block fixed but used randomly initialized feedback connectivity weights. The randomly initialized model performed poorly on the training data, but - depending on the strength of the feedback - does surprisingly well on the test data. The evaluation procedure used in the paper (we assign traces from the test data to the output channel of the model with the highest correlation), surely produces an upwards bias. Nevertheless, it seems that for some test conditions, already some unspecific feedback is sufficient, and the model structure contributes strongly to its overall performance. Please also note that in Fig. 4B,C, we are showing the Tonic Release Index (release under baseline), which is not simply explainable by a decrease in inhibition under the drug conditions.

Data (R1, R2 & R4) The training data are averaged over many animals/ROIs/repetitions, while the natural movie dataset consists of averages over only five repetitions and the sine dataset of single trial traces, making the two latter substantially more noisy. Furthermore, the datasets were collected under different experimental conditions, making it a harder generalization task because of the domain shift. We will stress this in the revised paper.

Model inference (R1) Fitting our model is a non-convex optimization problem. We tried to address this in Section 4.3-4.4 by using the 20 top performing models, for which we found consistent results. Exploring the whole parameter space would need different approaches (like posterior estimation) and is beyond the scope of this manuscript.

Connectomics (R1 & R4) For Section 4.3, we already compared our results to randomly sampled weight matrices. Fig. 9 shows a quantification of this comparison, and we will show a randomly generated, sorted matrix for illustration in the revised manuscript. The normalization of the rod BC and SAC connections was done because they can likely not be learned from a simple 1D stimulus, as they serve very specialized functions. This is an interesting direction for future research.

Frequency analysis (R4) We additionally analysed whether the BCN and the LSTM capture similar frequency ranges in the responses, using coherence on the generalization data as a measure. We did not find major systematic differences between the two models.