We appreciate the thoughtful feedback. All reviewers noted that our *sample validation* (SV) and *coordinated dropout* (CD) methods were novel with broad applicability. New analyses, clarifications, and proposed modifications are below.

R1: *Paper would be much stronger if ideas were demonstrated on multiple real datasets* Done (**Fig. 1a**). We used an open dataset [1] with a Random Target task (different lab and experiment). We found similar results to orig. Fig. 5, including the range where HP opt helps, and the gap between optimized and fixed HPs. **R1:** *Description of typical dataset sizes would help motivate the criticality of the issue*; *Single small dataset is insufficient to establish general efficacy.* Agreed, we'll discuss. Typical sizes largely vary, so for context we'll show the trial counts for 47 experiments from the open dataset (**Fig. 1b**; [1]). These dataset sizes are typical, and many are in the range where HP opt is important. Note: our original dataset (1836 trials) is actually *exceptionally large*, chosen so we could characterize HP opt vs. dataset size. **R1:** *Not clear why "Monkey J Maze" is not used from the beginning... Synthetic data is unconvincing.* This is a key point. It is important to clarify the necessity of tests on synthetic data, and may also help for readers without neural data experience.

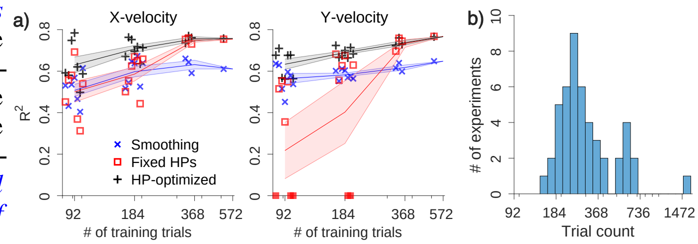

Fig. 1. **a**) Rand Targ task **b**) # of trials for 47 experiments [1]

The synthetic data is critical - without it, it is very challenging to determine whether an approach results in pathological overfitting. Real neural data has no ground truth for direct comparison - there is no "true", measurable firing rate. Common validation measures are problematic for detecting overfitting: *1)* Held-out likelihood of observed data is somewhat noisy and requires assumptions. *2)* Decoding behavior, as we do, is a rough measure: only a small fraction of neural activity correlates with behavior, and behavioral dynamics are quite slow. A precise characterization of overfitting (orig. Fig. 1) and of the effectiveness of SV/CD (orig. Fig. 4) would be very challenging with real data. Since SV & CD are the key innovations, we must thoroughly characterize them using data with a ground truth, and synthetic data are the best option. To speed manuscript, we will move all synthetic data generation details to a supplement. **R1:** *Existing regularization like denoising autoencoders (dAEs) should also be used as baselines. Motivation for completely new techniques should be explained.* Great suggestion. We tested dAEs (**Fig. 2**), and motivation is now easily explained in the context of these results. We repeated orig. Fig. 1b using two common dAE approaches for discrete data: 'Zero masking' and 'Salt and pepper noise' [2]. Important points: *1)* dAEs have a free parameter (noise level). *2)* Depending on its setting, dAEs can still show pathological overfitting. *3)* Some settings can even reduce performance. *4)* It is not possible to know how to set dAE noise *a priori*. Our methods bypass these limitations (see orig. Fig. 4), providing a reliable metric to measure (SV) or completely block (CD) pathological overfitting.

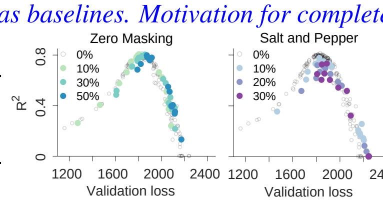

Fig. 2. Denoising AE results

R2: *Discuss if method can be extended to other data sets.* Good point, will add. Techniques should be applicable when forecasting time series from sparse data, especially when HP or architecture searches are important. Examples are usage at electrical vehicle charging stations, taxi/rideshare calls, etc.. We're currently trying to apply this to generative models for LIDAR/RADAR data for autonomous cars (e.g., following [3]). **R3:** *Would raise my score with the inclusion of some details that were missing... complete formulation of the generative model and inference procedure.* Good suggestion. We will add this information. R3's description of objective was accurate. **R3:** *State validation loss and how it is computed... Useful to fully describe LFADS model, at least in appendix*. Apologies for omissions, will add. **R3:** *Does the model still exhibit pathological overfitting with AR prior included?* Yes, and we were surprised by this (all the results in paper are with AR prior included). Key problem is AR prior is learnable, and model can adapt it to get better predictions by overfitting to spikes via inputs. Forcing a minimum AR prior autocorrelation might prevent overfitting, but might also prevent the model from capturing rapid changes. **R3:** *What HP settings provided "good" fits? Would be interesting to include a discussion, including how this might vary across dataset size.* Agreed, including settings/ranges will be helpful. Further, these methods enabled dynamic HP opt (changing HPs during training) using population based training [4]. This somewhat surprisingly yields even higher performance by learning schedules for different HPs (e.g., KL penalty is set high during early training, but decreases over time). We'll add this discussion. **R3:** *Is full-split CD necessary, or could you also split the data into input only, shared, and output only splits?* This is very interesting, we've been thinking about this also. The proposed 'Partial CD' approach might help when observed number of neurons is similar to the underlying dimensionality, and fully splitting data via CD may limit training. Without Full CD, though, a method is needed to detect/prevent overfitting. SV fills this role. As suggested, we turn CD on, and then allow some fraction of the data (searchable HP) to be shared as input and output. Preliminary tests on small sets of randomly drawn neurons (Monkey J Maze data, 25 per draw) show promising results: Partial CD outperforms Full CD in 8/10 models tested. Thorough tests will help delineate conditions where Partial CD helps.

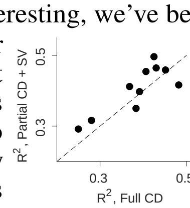

Fig. 3. Partial CD

[1] J E O'Doherty et al. http://doi.org/10.5281/zenodo.583331, 2017. [2] P Vincent et al. *J. Mach. Learn. Res.*, 11:3371–3408, 2010. [3] L Caccia et al. *arXiv:1812.01180*. [4] M Jaderberg et al. *arXiv:1711.09846*, 2017.