We thank all the Reviewers for their constructive comments. By addressing all the major concerns, we strongly feel that the paper has significantly improved. Reviewer #1 is positive and had some clarification questions. Reviewer #2 and #4 seemed to raise conceptually the opposite objections: Reviewer #2 questioned the value of non-linear neural data analysis methods in general, while Reviewer #4 was more concerned that we didn't compare our pi-VAE to technically more complicated methods. Interestingly, these two sets of comments do converge to the same practical issue, which is the lack of comparisons to the alternative methods, which we now address (see Fig. 1). We will first address these major concerns by Reviewers #2 and #4, then turn to minor issues raised by Reviewer #1 and others.

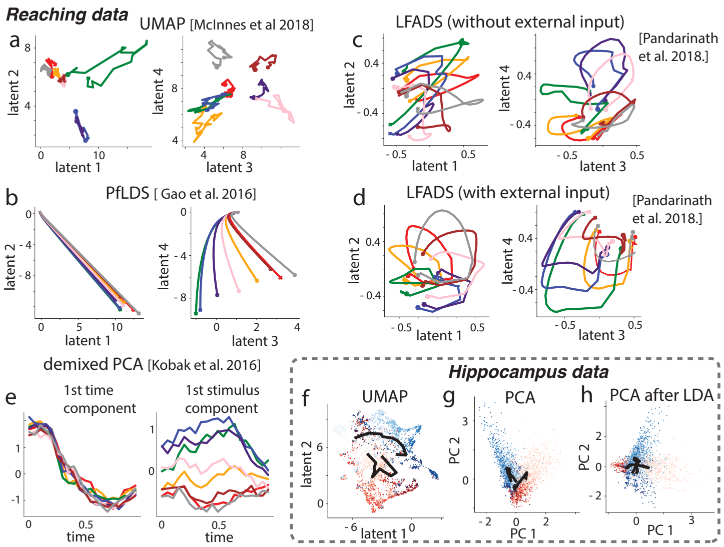

Figure 1: (a-e) Reaching data. (a-d) Color-coded, averaged latent trajectories corresponding to each reaching direction was plotted for each method. The filled dot and cross represent starting and ending of the trial. (e) The first set of components for demixed PCA. (f-h) Hippocampus data.

Most significantly, we have analyzed the data using the methods suggested by the Reviewers, including linear (demixed PCA or dPCA, PCA, LDA+PCA) and nonlinear methods (supervised UMAP, PfLDS, LFADS). Whenever possible, the original authors' implementations of the methods were used. Fig. 1 summarizes some of these new results. **Reaching data:** Overall we find that, while the extracted latent structures from these methods exhibit interesting characteristics, none of them results in fully disentangled latents. Supervised UMAP recovers different directions as different clusters, but without clear representations of temporal dynamics. LFADS and PfLDS both lead to smooth trajectories. Although the trajectories for different directions are separated in the 4-dimensional space, directions and temporal dynamics are entangled so that it is difficult to interpret each individual latent dimension (Fig. 1b,c,d). dPCA with both time and directions as labels still entangles time and directions (stimulus components change with time) (Fig. 1e). **Method considerations:** 1. LFADS can take task variables as external inputs to the model RNN. We thus try LFADS with or without reaching direction as external inputs (Fig. 1c,d). 2. dPCA only deals with discrete task variables each with the same number of trials and each trial with the same length, and is unable to recover additional latent fluctuations as our method. 3. UMAP can incorporate label information for supervised learning, and we use the reaching directions as labels (Fig. 1a). However, it doesn't recover temporal dynamics. **Take-away:** In latent space recovered by pi-VAE, dimensions separately encode temporal dynamics and reaching direction (as shown in our paper). In contrast, for latents learned by the alternatives, they are entangled and cannot be easily decoupled even after rotations (Fig. 1). Similar entangled results are found when applying these alternatives with 3-dimensional latents. Additionally, we find that pi-VAE could recover the geometry of the physical reaching targets, while other alternatives cannot. **Hippocampus data:** Applying the suggested methods (where supervised UMAP takes locations as labels), we find that the resulting latents are all less interpretable than pi-VAE (Fig. 1f,g,h), with no dimension directly representing the rat's location. Also the rhythmic-like fluctuations span across dimensions, rather than concentrate in one dimension (not shown). **Revision:** We will incorporate these plots and add detailed discussions on comparisons to the alternatives in revision, along with relevant references. These new results/comparisons substantially strengthen our conclusions, and they further highlight the identifiability and interpretability of latent representation recovered by pi-VAE. **Code availability**: Please be assured that the code will be documented and made available upon publication.

We would like to further note that these methods have different motivations and focuses, and how well they work in practice may depend on the questions being studied. Our method is motivated by leveraging the strength of regression-based methods and latent models to increase the identifiability and interpretability, a direction received little attention previously. To reduce the number of assumptions, we did not incorporate temporal smoothness priors, which are key for PfLDS, LFADS. Probably it would be best to consider these methods as complementary rather than competing methods. These points were briefly mentioned in the Discussion, and we will revise the text to make them more explicitly.

For questions raised by Reviewer #1: 1. Yes, we meant "inferred without label prior". 2. Yes, Poisson noise was used for VAE, and the only difference is that label prior was not used. 3. We completely agree that more emphasis should be put on difference between Fig. 3f and 3j, Fig. 4b and 4d. Thank you for this excellent suggestion. 4. The statement of "Poisson noise, continuous labels..." was made in reference to the specific implementation in the GIN paper. We agree that it was misleading. Finally, space limit prevents us from a discussion of the advantages of GIN here, please refer to their original paper. We will fix all these points, along with other minor concerns in the revised version.