We are very grateful for the thoughtful feedback provided by the reviewers. We will incorporate all minor comments, including the suggestions on notation given by reviewer 2 and the comments by reviewer 3 regarding least-squares sub-optimality and clarity regarding the Gibbs sampler. We respond the main substantive comments below.

R1 *"The biggest drawback seems to be that the model is over parametrized... and therefore there is a possibility of finding spurious motifs... Experiments determining whether there is overfit should be done."*

Overfitting is certainly a possibility. Our approach is to generalize the cross-validation procedure used for convNMF (Mackevicius et al., 2019), which is related to a well-established procedure for PCA ("speckled holdout"; Wold, 1978). We evaluate the predictive likelihood after each MCMC sample, thus building an approximation to the full predictive posterior (see fig. 2). This agrees with standard Bayesian cross-validation practices (see, e.g., Ch. 7 of Gelman et al.).

To avoid any potential confusion, we note that our model is not "over-parameterized" in the typical sense unless the number of motif types, R, is larger than the number of neurons, N. We do not consider this regime, though it might be of interest—it could be viewed as a point process analogue to sparse dictionary learning.

R1 *"By definition, a motif is a motif only when it recurs in the data (otherwise anything could vacuously be a motif)."*

We agree that such vacuous motifs might be identified during the training phase, but they would not lead to above-chance performance in the test set. In practice, we do not expect this to be a concern—neuroscientists will typically use pp-Seq on very long time series and with hyperparameter settings that render such vacuous motifs extremely unlikely.

R1 *"Using an AIC/BIC like criterion would help [choosing the number of motif types]."*

Due to the non-parametric nature of our model (the number of latent events K is random), is not immediately obvious to us how to compute the correction factors in AIC and BIC. These criteria are typically applied when identifying point estimates of parameters (e.g. in maximum likelihood inference), while we build an approximation to the full posterior. Nonetheless, at a high level, AIC and BIC aim to approximate something similar to cross-validation. Computational expense is the only downside we see to cross-validation, and in practice we have not found this to be prohibitive.

R2 *"The choice of priors... may be tricky when applying this method to a different dataset"*

We expect practitioners to choose priors that reflect the appropriate order-of-magnitude of expected sequences—e.g. for the rat dataset the prior for sequence length was ~ 10s while for songbird data we chose ~ 1s. These choices were guided by our domain knowledge, which future practitioners will also draw upon. In many interesting biological datasets, the same neural sequences reoccur many times, suggesting that a suitably weak prior will be dominated by the likelihood term, and so the model should produce the desired result. Using synthetic data, we have observed that a weak, but misspecified, prior recovers the ground truth sequences—we will look for a way to incorporate this into our revised paper. Note that practitioners could also use cross-validation to compare amongst different hyperparameter choices.

R3 *"One lingering question was how well does the Gibbs sampler converge?... If different realizations of the chain end up giving somewhat different results... what recommendations can the authors provide..?"*

In all of our experiments, the sampler converged to very similar parameter ranges. Fig 3B shows, e.g., that the number of sequence events, K, is very similar across three independent MCMC runs. Similar results hold for other parameters; we will add additional details in our revision. This concern is not unique to our model—slow mixing across separated modes is a well-known problem for MCMC inference with no easy solution. ConvNMF can also converge to different solutions across optimization runs. In practice, neural sequences are often salient enough to overcome these worst-case conclusions. We will also look into additional MCMC convergence diagnostics (e.g. Gelman-Rubin) for our revision.

R4 *"I suppose that the number of types R and number of events K are hyperparameters that require predetermined before inference. How were these numbers determined in this study?"*

The number of sequence types R must be specified, along with a few other hyperparameters (we propose to use cross-validation to guide these choices, as discussed above). The number of events, K, *does not need to be specified*—since this changes over MCMC samples we obtain an approximation to the posterior—i.e., $p(K \mid \texttt{data})$ informally.

R4 *"The global parameter Theta however does not benefit from the parallel execution. How was the parameters (including the time warping parameter) learned?"*

Our current implementation does not use parallel computation in the global update since this step is very fast and not the primary bottleneck (in principle, some of these computations could be further parallelized). The Gibbs update over a grid of W warping values is analogous to re-sampling the motif type over R possibilities. Due to space constraints, the details are in Supplement E, but we will make an effort to provide more intuition in the main text in our revision.