We thank the reviewers for their thoughtful comments. We appreciate that all three reviewers affirmed the importance of our work and the rigor of our approach. We believe that no major weaknesses were identified by the reviews. In our view, the comparisons between recurrent neural network models and experimental data are one of the most important contributions of our work, and all reviewers agreed that this was a core strength of the manuscript.

The reviewers highlighted several future modeling directions that are raised by our results and that we did not explore in the manuscript. For example, Reviewer 2 suggests that we train networks on a navigation task alone, freeze the weights, and then train on a context discrimination task. We agree that this kind of contextual learning paradigm is of interest and could provide insight into biological remapping, such as that observed by Low et al. (2021). We also agree with Reviewer 3's broader point that "There are many choices that must be made when simulating RNNs and there is a growing awareness that these choices can influence the kinds of solutions RNNs develop." It is notable that we were able to reproduce the qualitative features of the experimental data without finely tuning hyperparameters (we used default settings in PyTorch layers), using a very basic training protocol (gradient descent with gradient clipping), and without adding any hand crafted regularization (though we agree that regularization could make the RNN solution look even more like the data).

We believe that readers will benefit from reading the reviewers' suggestions, which are insightful and well-motivated. Having weighed the reviewer comments carefully, we feel that our manuscript stands as a complete scientific story. We hope that the public reviewer comments will inspire future investigations to fully explore these possibilities and unpack their outcomes at a level of detail that would not be possible in the context of our manuscript.

Thus, we have chosen to implement the following minor changes suggested by the reviewers, which we hope will improve the clarity of the text and figures (summarized below). These changes do not alter the fundamental content of the manuscript.

Text:

(Where appropriate, strikethrough indicates old text, bold indicates new text.)

- We corrected a few minor typos.
- We updated the citations to follow the eLife citation style.
- To address comments from Reviewers 1 and 2: we reworded the final paragraph of the Introduction (p. 3) to remove the term "modularity" and clarify our main finding. Those sentences now read, "The RNN geometry and algorithmic principles readily generalized from a simple task to more complex settings. Furthermore, we performed a new analysis of experimental data published in Low et al.[26] and found a similar **geometric** ~~modular~~ structure in neural activity from a subset of sessions with more than two stable spatial maps."
- To address comments from Reviewer 1: in the first paragraph of the Results section *A recurrent neural network model of 1D navigation and context inference remaps between aligned ring manifolds* (p. 3), we added the sentence, "Remapping was not aligned to particular track positions, rewards, or landmarks." to clarify that experimental result from Low et al. (2021).

- To address comments from Reviewer 3: in the final paragraph of the Results section *Aligned toroidal manifolds emerge in a 2D generalization of the task* (p. 11) we clarified that models were trained "to estimate ~~2D~~ position **on a 2D circular track.**" We also added a citation to Cueva, Ardalan et al. (2021) with the following sentence, "Notably, each toroidal manifold alone is reminiscent of networks trained to store two circular variables without remapping."
- To address a question from Reviewer 2: in the final paragraph of the Results section *Manifold alignment generalizes to three or more maps* (p. 13), we added the following clarification: "In Supplemental Figure 3, we show that RNNs are capable of solving this task with larger numbers of latent states (more than three**; for simplicity, we consider up to 10 states**)."
- To address a comment from Reviewer 1: in the fourth paragraph of the Discussion (p. 17), we removed the sentence, "Notably, our model captured aspects of the data that these previous forward-engineered models did not explore—namely, that the ring manifolds corresponding to the correlated spatial maps were much more aligned than expected by chance and than strictly required by the task." to focus on the key point in the following sentence that, "forward-engineered models provide insights into *how* neural circuits may remap, but do not answer *why* they do so."
- To address comments from Reviewers 1 and 2: we reworded the penultimate paragraph of the Discussion (p. 17–18) to clarify our findings and remove the term "modularity" (except when referencing papers that themselves use that term (Driscoll et al., 2022; Yang et al., 2019)). Those sentences now read:

  When ~~the~~ RNN architecture is explicitly ~~modular in its design~~ **designed to include dedicated neural subpopulations**, these subpopulations can improve model performance on particular types of tasks (Beiran et al., 2021; Dubreuil et al., 2022). Thus, there is an emerging conclusion that RNNs use simple dynamical motifs as building blocks for more general and complex computations, which our results support. In particular, aligned ring attractors are a recurring, ~~modular~~ **dynamical** motif in our results, appearing first in a simple task setting (2 maps of a 1D environment) and subsequently as a component of RNN dynamics in more complex settings (e.g., as sub-manifolds of toroidal attractors in a 2D environment, see Figure 4). We can therefore conceptualize a pair of aligned ring manifolds as a dynamical "building block" that RNNs utilize to solve higher-dimensional generalizations of the task. Intriguingly, ~~this modular motif also emerged in~~ our novel analysis of neural data from Low et al. (2021)~~, suggesting that biological systems may also leverage this strategy~~ **revealed that similar principles may hold in biological circuits—when three or more spatial maps were present in a recording, the pairs of ring manifolds tended to be aligned**.

- To address questions from Reviewers 2 and 3: in the first paragraph of the Methods section *RNN Model and Training Procedure* (p. 21), we added the sentence: "The connection weights were randomly initialized from the uniform distribution $U(-\sqrt{\frac{1}{N}}, \sqrt{\frac{1}{N}})$, which is the default initialization scheme in PyTorch."
- To address a question from Reviewer 2: we added a third paragraph to the Methods section *Manifold Geometry Analysis* (p. 23), as follows:

In Figure 1K, 4G, 5G, and Supplementary Figure 2B, we calculate the angles between the input and output weights and the position subspace or remapping dimension. To find this angle, we calculated the cosine similarity between each weight vector and each subspace. Cosine similarity of 0 indicates that the weights were orthogonal to the subspace, while a similarity of 1 indicates that the weight vector was contained within the subspace.

- To address a question from Reviewer 1: we added the following sentence to the second paragraph of the Methods section *Experimental Data* (p. 24), "We performed the same analysis of trial-by-trial spatial stability to obtain the similarity matrices in Figure 1C and G."

Figures and legends:

- To address a question from Reviewer 1: in Figure 1C and G, we added x-axis labels to the similarity matrices to clarify that these are trial-by-trial correlations.
- To address a question from Reviewer 1: we expanded the Figure 1C legend to clarify the experimental results as follows:

  Old legend:
  (C, left) An example medial entorhinal cortex neuron switches between two maps of the same track (top, raster; bottom, average firing rate by position; red, map 1; black, map 2). (C, right/top) Network-wide trial-by-trial correlations for the spatial firing pattern of all co-recorded neurons in the same example session (colorbar indicates correlation). (C, right/bottom) k-means map assignment.

  New legend:
  (C, left) An example medial entorhinal cortex neuron switches between two maps of the same track (top, spikes by trial and track position; bottom, average firing rate by position across trials from each map; red, map 1; black, map 2). (C, right/top) Correlation between the spatial firing patterns of all co-recorded neurons for each pair of trials in the same example session (dark gray, high correlation; light gray, low correlation). The population-wide activity is alternating between two stable maps across blocks of trials. (C, right/bottom) K-means clustering of spatial firing patterns results in a map assignment for each trial.

- To address comments from Reviewer 3: in the legend of Figure 4C, we added the sentence "Note that the true tori are not linearly embeddable in 3 dimensions, so this projection is an approximation of the true torus structure."
- To address a question from Reviewer 2: we expanded the legend for Supplementary Figure 2 to clarify the purpose of the figure schematics as follows:

  Old legend:
  (A) Schematic showing the orthogonalization of the position and context input and output weights.
  (B) Reproduced from Figure 1K.
  (C-D) Schematic: How a single velocity input (blue arrows) updates the position estimate (yellow to red points) from the starting position (blue points).
  (C) Velocity input lies in the position tuning subspace (gray plane)(hypothetical). Note that the same velocity input results in different final positions.

(D) Velocity input is orthogonal to the position tuning subspace (observed).
(E) Schematic of possible flow fields in each of the three planes (numbers correspond to planes in C and D), which would result in the correct positional estimate given orthogonal velocity inputs at different positions (D).

New legend:
(A) Schematic showing the relative orientation of the position output weights and the context input and output weights to the position and state tuning subspaces.
(B) Reproduced from Figure 1K.
(C-D) Schematic to interpret why the position input weights are orthogonal to the position tuning subspace. These schematics illustrate how a single velocity input (blue arrows) updates the position estimate (yellow to red points) from a given starting position (blue points).
(C, not observed) Velocity input lies in the position tuning subspace (gray plane). Note that the same velocity input pushes the network clockwise or counterclockwise along the ring depending on the circular position
(D, observed) Velocity input is orthogonal to the position tuning subspace and pushes neural activity out of the subspace.
(E) Schematic of possible flow fields in each of three planes (numbers correspond to planes in C and D). We conjecture that these dynamics would enable a given orthogonal velocity input to nonlinearly update the position estimate, resulting in the correct translation around the ring regardless of starting position (as in D).