Hippocampal attractor dynamics predict memory-based decision making

Ben Steemers[1,2], Alejandro Vicente-Grabovetsky[1], Caswell Barry[3], Peter Smulders[1], Tobias Navarro Schröder[1], Neil Burgess[4,5] and Christian F. Doeller[1]

[1]Donders Institute for Brain, Cognition and Behaviour, Radboud University, 6525 EN Nijmegen, the Netherlands

[2]Laboratory of Neuropsychology, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892, USA

[3]Research Department of Cell and Developmental Biology, UCL, Gower Street, London WC1E 6BT, UK

[4]UCL Institute of Cognitive Neuroscience, 17 Queen Square, London WC1N 3AZ, UK

[5]UCL Institute of Neurology, Queen Square, London WC1 3BG, UK

Contact

Correspondence should be addressed to Ben Steemers (ben.steemers@nih.gov) or Christian Doeller (christian.doeller@donders.ru.nl).

Summary

Memories are thought to be retrieved by attractor dynamics if a given input is sufficiently similar to a stored attractor state [1-5]. The hippocampus, a region crucial for spatial navigation [6-12] and episodic memory [13-18] has been associated with attractor-based computations [5; 9], receiving support from the way rodent place cells 'remap' nonlinearly between spatial representations [19-22]. In humans, nonlinear response patterns have been reported in perceptual categorization tasks [23-25], however, it remains elusive whether human memory retrieval is driven by attractor dynamics and what neural mechanisms might underpin them. To test this we used a virtual reality [7; 11; 26-28] task where participants learned object-location associations within two distinct virtual reality environments. Participants were subsequently exposed to four novel intermediate environments, generated by linearly morphing the background landscapes of the familiar environments, while tracking fMRI activity. We show that linear changes in environmental context cause linear changes in activity patterns in sensory cortex, but cause dynamic, non-linear, changes in both hippocampal activity pattern and remembered locations. Furthermore, the sigmoidal response in the hippocampus scaled with the strength of the sigmoidal pattern in spatial memory. These results indicate that mnemonic decisions in an ambiguous novel context relate to putative attractor dynamics in the hippocampus, which support the dynamic remapping of memories.

Results

To create stable object-place memories, we let participants extensively learn the locations of four objects in two virtual environments (environment A and F; Figure 1) over the period of two days, while feedback about the correct object position was provided at the end of each trial. Performance was measured as the distance error in object replacement as a fraction of arena width. Throughout the two training sessions, participants' performance increased, reaching ceiling levels on day 2 (Figure S1A).

Performance did not differ between the two environments at the end of training (t_{19}=1.19, P=0.25), see Figure S1.

After training, and while lying in the MRI scanner, participants were required to perform the same behavioral task in four novel 'morph' environments (B through E) in addition to the known environments A and F. Crucially, the backgrounds, which distinguished the environments, varied linearly from A to F (e.g. B = 80%*A + 20%*F, C = 60%*A + 40%*F, etc., Figure 1B). Unknown to the participants, transitions between environments were introduced during inter-trial intervals. Participants were not informed about the environmental manipulation and did not receive feedback during this session. Environments were presented in a random order (Supplemental Experimental Procedures).

To formally assess the behavioral response profile, we looked at the relative difference between the object replacement location and the true object location in environments A (ΔA) and F (ΔF), expressed as ΔA/(ΔA+ΔF). This behavioral similarity measure scales linearly from 0 to 1 with increasing ΔA and decreasing ΔF, reflecting the 'A-ness' and 'F-ness' of each memory response (Figure 1A). We used maximum likelihood estimates (MLE) to fit this measure to (I) a sigmoidal model indicative of putative attractor dynamics, and (II) a linear control model representing the visual change in environments A to F, with model complexity held constant (both models have 2 free parameters; Supplemental Experimental Procedures). The similarity measure indeed followed a sigmoidal rather than a linear model (paired t-test on the resulting residual sum of squares (RSS), t_{19} = 4.62, P < 0.001; Figure 1D), with the sigmoid centered between environments C and D in the majority of participants (15/20; Figure 1E). On the other hand, memory confidence, measured on a 5-point scale after each trial, scaled linearly with difference from either baseline environment in 87% of participants (Figure 1D). A behavioral control experiment, where naive participants judged the similarity between the background

cues, showed that the sigmoidal memory response pattern was not due to differences in the perception of the backgrounds, which were judged to differ in a linear rather than a sigmoidal fashion along the morph sequence (t_{15}=4.61, P0.001; Figure 2).

If the sigmoidal behavioral pattern reflects putative attractor dynamics in brain activity, we would expect a similar sigmoidal transition in the similarity of the multi-voxel activity patterns for the morph environments compared to the two baseline environments [29]. Since the offset of the sigmoids fitted to behavioral data differed between participants (see Figure 1E), we used each participants' behavioral response pattern to predict the similarities in multi-voxel activity patterns between pairs of environments, akin to population vector analyses of place cell firing [20] (Supplemental Experimental Procedures). We correlated the actual multi-voxel patterns for each environment-by-environment combination and tested this against the aforementioned prediction models using general linear modeling (GLM) (Figure 3A). We found a response pattern following the sigmoidal prediction obtained from each participant's behavioral response function in a hippocampal ROI (peak coordinates x=-31, y=-26, z=-7, peak Z=3.10 uncorrected; bootstrap corrected P=0.036, see Figure 3B and Supplemental Experimental Procedures for details). The sigmoidal effect in the hippocampus was strongest for responses with high memory confidence (Figure 3D). No effect was found in the hippocampus for a linear model, even at a lenient uncorrected threshold of P<0.01. In contrast, a linear, but not a sigmoidal response pattern was observed in visual cortex (x=-8, y=-79, z=11; peak Z=3.91 uncorrected; bootstrap corrected P<0.01, Figure 3C; no region outside visual cortex showed a significant linear effect at bootstrap-corrected P<0.05). Additional control analyses demonstrated that the sigmoidal effect in the hippocampus was not due to differences in navigational behavior across environments (Figure S2A and B), differences in mean hippocampal BOLD signal across environments (Figure S2C), nor an effect of the similarities of behavioral response trajectories when the same object was placed across environments (Figure S3A). Finally, a consistent sigmoidal or linear response pattern

was absent in the perirhinal cortex, entorhinal cortex and parahippocampal cortex (Figure S3B), making it unlikely that our hippocampal observation is the net result of putative extrahippocampal attractor dynamics within the medial temporal lobe.

In sum, these results indicate that the neural activity pattern in the hippocampus follows nonlinear dynamics matching the behavioral response pattern. However, the presence of a sigmoidal response pattern in brain and behavioral data does not unambiguously imply that both adhere to putative attractor dynamics to a corresponding degree. To test this, we assessed the strength of the sigmoidal response profile in behavioral and fMRI data separately by fitting a 'perfect', canonical, sigmoidal model using GLM to both data-sets (i.e. a step function predicting immediate transition between A and F states, see Figure S4A). The resulting t-values for both types of data, obtained per participant, were correlated across participants (Figure S4B). Significant correlation was seen again in the hippocampus (peak coordinates x=-29, y=-15, z=-19; peak Z=3.79 uncorrected; bootstrap corrected P=0.002; Figure 4A and B) for the perfect sigmoidal model but the effect was absent for the linear model (even at a lenient threshold of P<0.01, uncorrected). Importantly, an additional within-trial analysis showed that the ratio between linear and sigmoidal fit systematically changed between the early and the late phases of trials, reflecting a dynamic shift to a dominantly sigmoidal fit towards late trial phases (linear regression: slope=0.004, P<0.005; Figure 4C). Furthermore, Monte-Carlo simulations showed that the linear model outperforms the sigmoidal model on randomly shuffled data (Kolmogorov-Smirnov test: P<0.001, Figure 4D; Supplemental Experimental Procedures). Finally, additional post hoc analyses further suggest that the behavioral and fMRI responses are both clustered around environment A or F representations (Figure 4F), indicative of a concurrent sigmoidal pattern in behavior and neural data.

Discussion

Our data show that memory retrieval in ambiguous novel situations is associated with nonlinear

dynamics in the hippocampus, the brain's key region for episodic and spatial memory [7; 9; 13-18; 28-30], corroborating predictions of attractor-based computational models of memory [1-4]. Our findings also dovetail with recording work in the hippocampus [20; 22] by demonstrating a nonlinear response pattern in the hippocampus as a function of linearly changing input. Furthermore, the sigmoidal response pattern in the hippocampus was (I) predominant in trials in which participants had high confidence in their memory response and (II) scaled with the strength of the sigmoidal pattern in behavior across participants. Thus, the current findings shed new light on the behavioral relevance of these hippocampal computations. That is, we demonstrate that the divergence of orthogonal, competing, representations in the hippocampus directly translates into mnemonic decisions, indicative of putative attractor dynamics [13; 17-18; 31-35].

The putative neural process underlying the formation of hippocampal memories is remapping [19], the formation of distinct representations by populations of place cells in response to environmental change. Place-cell-based representations exhibit attractor-like dynamics (sharp transitions) when animals are exposed to similar novel environments that have features mapping in-between already known, distinct, environments (differing in shape, color and texture; [22]), but not if trained in environments that are less distinct [36]. Although it is infeasible to register place cell activity in humans using fMRI, given that the distribution of place cells in rodent hippocampus is non-topographic with respect to the spatial distribution of their firing fields [37], the aforementioned rodent place cell pattern shows striking similarities with our data. Our data are consistent with the absence of sigmoidal neural response patterns in the human hippocampus when participants view highly similar visual scenes [38-39] and a linear scaling of hippocampal responses with changes in the configuration of landmarks in four virtual environments [40]. The attractor-like behavior of place cells also accords with observations in rodents that repeated exposure to less distinct environments is accompanied by a slower, gradual development of distinct place cell representations [41]. However, the distribution of the place cell activity in

different environments has not been examined in detail, and it is the similarity of activity in different environments that we examine with MVPA in our study. Our results might also relate to a recent observation by Agarwal et al. [42], showing that spatial information can be obtained from local field potentials (LFP) in the rodent hippocampus, summating electrical activity of a large number of neurons, which more closely relates to the BOLD response obtained with fMRI. However, future translational studies are necessary to more directly understand the relationship between population activity of spatially tuned cells and the fMRI signal in the hippocampal formation.

How do nonlinear dynamics in the hippocampus relate to pattern separation, as for instance indexed by fMRI adaptation [43]? Pattern separation, albeit related, is not necessarily due to attractor dynamics. Pattern separation in the hippocampus might accentuate small differences, rather than being attracted to a different fixed point, like the familiar pattern of the baseline environments in the present study and in the work by Wills et al. [22]. In addition, although pattern separation has been observed during virtual reality navigation [44], putative attractor dynamics are characterized by the emergence of a nonlinear response profile over time (Figure 4C).

Could the presence of a sigmoidal effect in the fMRI pattern in the hippocampus be explained by something other than putative attractor dynamics? A nonlinear response profile as observed in the present study effectively reflects a thresholded tuning function. Here we provide evidence for a sigmoidal pattern in both behavioral and multi-voxel fMRI data as well as in a brain-behavior interaction and also show the systematic emergence of the nonlinear response profile within trials. Taken together, these effects are indicative of putative attractor dynamics, but ultimately not a necessary condition and alternative processes could potentially explain the sigmoidal output function. Any such alternative must either include the differences in background images, or the locations where objects are placed by participants, since these are the only features that change between environments.

Eliminating same-object comparisons in our analysis does not influence our results, therefore the drop locations and associated similarities of paths and background cues cannot drive the observed sigmoidal pattern (Figure S3A). Furthermore, in visual cortex we do see a linear response pattern rather than a sigmoidal pattern, making it further unlikely that mere visual differences between environments contribute to the observed effect (Figure 3C). In addition, naive participants perceive the background images to change linearly (Figure 2) and this perception could be used to make a binary choice on where to drop the cued object (i.e. if perceived more similar to A, recall environment A). This strategy would eliminate the need for putative attractor dynamics in the hippocampus and still show binary memory activation in fMRI data. However, it would not predict a change from a more linear to a more sigmoidal pattern within trials as seen in our data (see Figure 4C) nor be congruent with earlier rodent studies [22].

In sum, the sigmoidal response pattern we observe in the hippocampus provides novel evidence for an abrupt, remapping-like response to a linearly changed spatial context in humans, consistent with a recent report showing consequences of such a response in multi-modal pattern completion [45] and with pattern separation/completion during memory disambiguation in virtual reality [44]. Participants were trained to ceiling so as to form distinct and separate memories for the two original environments. In addition, analyses of fMRI data acquired during the learning phase for a subset of our participants indicates that, like place cell remapping in less distinct environments [41], representations of environments A and F in the hippocampus de-correlate as a function of learning such that distinct representations already emerge on day 1 and continue to diverge on day 2 (Figure S1; see Supplemental Experimental Procedures). This suggests that a certain level of distinctiveness between two representations is required to observe nonlinear dynamics between the morph environments.

In conclusion, our study provides evidence for putative attractor dynamics and spatial remapping in the human hippocampus, and highlights that these neural mechanisms underpin memory-based decision making in novel situations.

Author Contributions

Christian Doeller (CD), Neil Burgess (NB) and Caswell Barry (CB) conceived this research; CD, CB, Ben Steemers (BS) and Alejandro Vicente-Grabovetsky (AVG) designed the experiment. BS, Tobias Navarro Schröder (TNS) and AVG developed the experimental task and performed the fMRI experiment. Peter Smulders (PS) performed the behavioral experiment. BS, AVG and CD analyzed the data. CD, BS, TNS, NB, CB and AVG wrote the paper. All authors discussed the results and contributed to the paper.

Acknowledgments

This work was supported by European Research Council (ERC-StG 261177) and Netherlands Organization for Scientific Research (NWO-Vidi 452-12-009) fellowships awarded to CD. NB is supported by the Wellcome Trust and the Medical Research Council, UK. CB is funded by the Wellcome Trust and the Royal Society, UK (101208/z/13/z). The authors would like to thank S. Auger and O. Vikbladh for earlier behavioral pilot work, S. Bosch for help with data acquisition and A. Backus and J. Bellmund for useful discussions. Correspondence should be addressed to Ben Steemers (ben.steemers@nih.gov) or Christian Doeller (christian.doeller@donders.ru.nl).

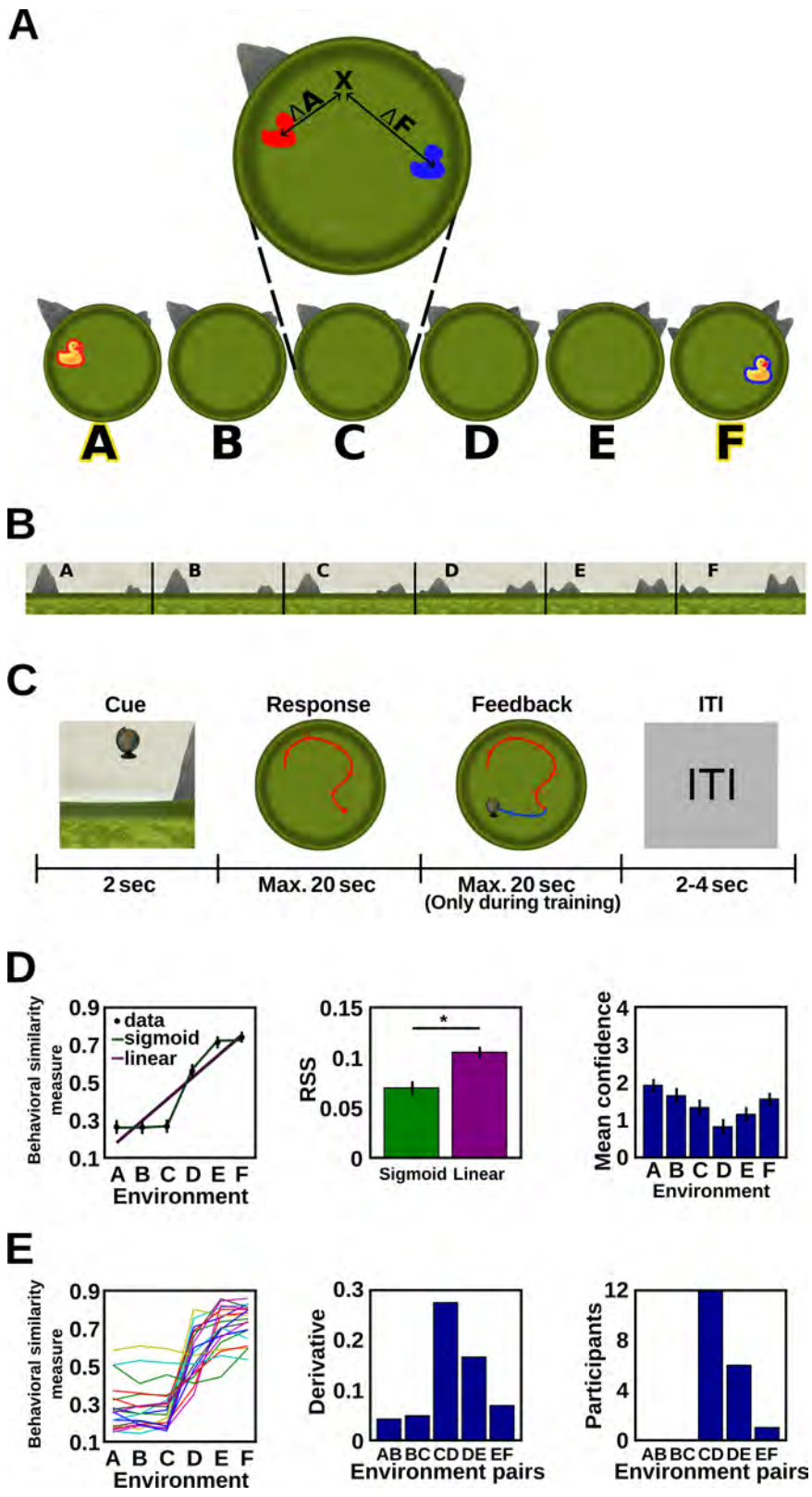

Figure 1. Memory performance in the virtual reality task. (**A**) Participants learned the locations of four objects in two environments (A and F), the duck depicts an example object. Subsequently, they

‘replaced’ the objects in environments A and F, and in morphed environments B, C, D and E. For every trial the distance of replacement locations (illustrated by ‘X’ in the example of environment C) from the object’s positions in A (ΔA, see red duck) and F (ΔF, see blue duck) was measured. (**B**) Morph sequence of background cues. Morphing A into F was achieved by changing the contribution of the height maps of the mountains in the background from environment A to F in a linear fashion. (**C**) In each trial participants were cued with one of the objects, then navigated to the remembered object location and placed the object by a button press (response). Feedback was given by showing the object in its correct location (only in the training phase). (**D**) Left: the relative sizes of ΔA and ΔF (see **A**), expressed as ΔA/(ΔA + ΔF), was taken as a behavioral expression of the similarity of any environment to the base environments A and F. This behavioral similarity measure is plotted separately for the different environments (averaged across trials and participants, +/- SEM) along with sigmoid and linear model fit curves. Middle: bars show model fits (residual sum of squares, RSS, between model and data, +/- SEM) separately for both models. The sigmoid model fits the data better than the linear model (t_{19}=4.62, P<0.001). Right: Mean confidence (averaged across participants +/- SEM) is plotted for each environment. ANOVA shows a significant effect of environment on confidence ($F_{(5,70)}$=10.92, P<0.001). (**E**) Left: the relative difference between the drop error in environment A (ΔA) and F (ΔF), expressed as ΔA/(ΔA+ΔF), is plotted per environment (averaged across trials), separately for each participant. Middle: mean derivative between the responses from panel **D** between subsequent environments are plotted. The derivatives differ significantly ($F_{(4,76)}$=53.6, P<0.001) and peaks between environment C and D. Right: the highest average slope of linear fits on ΔA/(ΔA+ΔF) between neighboring environments is observed in most participants between environment C and D. (For performance during the training phase see Figure S1).

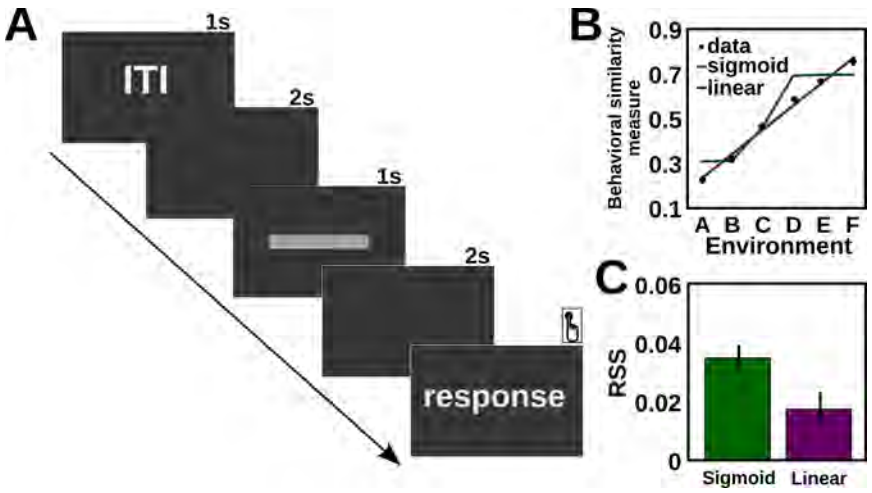

Figure 2. Perception of environment-specific background independent of virtual navigation. (**A**) Trial structure of follow-up behavioral experiment. 16 naive participants scored the similarity between pairs of background images on a 5-point scale. These were the same images used to make distinguishable backgrounds in the 3D navigation task but now presented 2D. (**B**) Naive similarity judgments for each environment to either environment A (ΔA) or environment F (ΔF), expressed as ΔA/(ΔA+ΔF), are plotted, along with linear and sigmoidal model fits (averaged across participants +/- SEM). (**C**) The measure ΔA/(ΔA+ΔF) was fitted to a linear and perfect sigmoid model using MLE and the resulting RSS are shown (+/- SEM). Comparison of the residuals revealed that the linear model shows a significantly better fit than the sigmoidal model (paired t-test t_{15} =4.61, P<0.001).

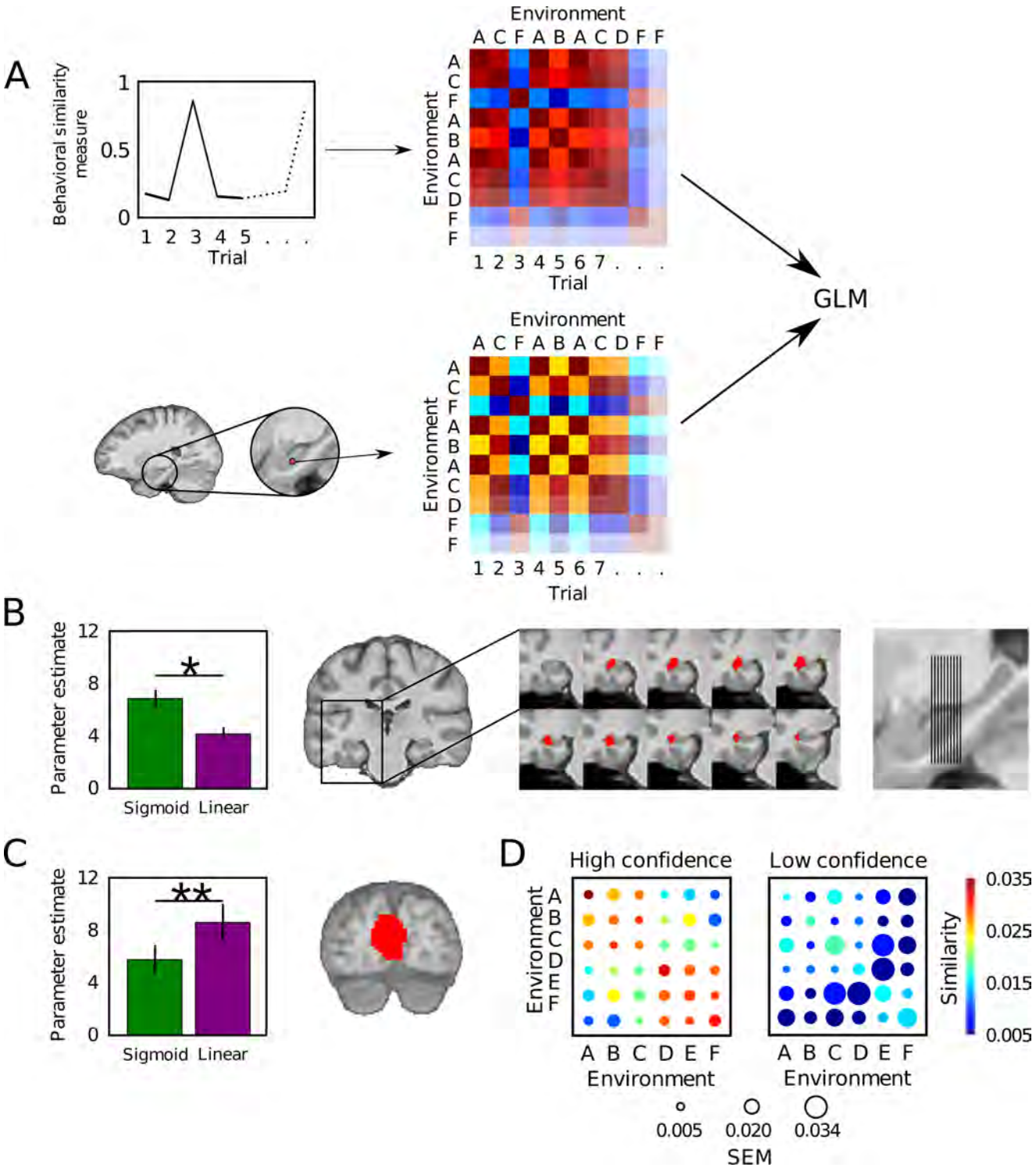

Figure 3. Sigmoidal response pattern in the hippocampus. (**A**) Schematic of analysis logic: Behavioral similarity to baseline environments (see example in top left panel) were used to generate predictor matrices for all environment pairs (top right). Multi-voxel patterns in a searchlight analysis

(bottom left, see Supplemental Experimental Procedures) were correlated between trials creating a data matrix (bottom right), which was tested against the predictor matrix using GLM (see Figure S4A). (**B**) Results from the behaviorally-informed sigmoidal and the linear prediction model restricted to hippocampal region-of-interest, bars show the average effect size in the hippocampus peak +/- SEM (x=-31, y=-26, z=-7; bootstrap corrected $P<0.05$; see also Figure S3). Results, thresholded at a bootstrap corrected $P<0.05$, are overlaid on a study-specific structural template and resampled to MNI space. Depicted is the extent of the hippocampal effect from y=-34 to y=-24 with slice locations shown on a sagittal plane (right). These results were not influenced by differences in navigational behavior or mean hippocampal BOLD signal across environments (see Figure S2). (**C**) Searchlight results using a linear prediction model are overlaid on an averaged structural image (y=-85), thresholded at bootstrap corrected $P<0.05$. Bars show the effect size in the peak visual cortex voxel +/- SEM (x=-8, y=-79, z=11; bootstrap corrected $P<0.005$), separately for the sigmoidal and linear model. (**D**) Strength of neural similarity of hippocampal multi-voxel pattern for any environment compared to all other environments (radius of 5mm around the peak from the sigmoid model, see **B**; averaged across participants). Color code reflects strength of neural similarity, size of circle indicates SEM. Left plot refers to high confidence trials, right plot shows data for low confidence trials.

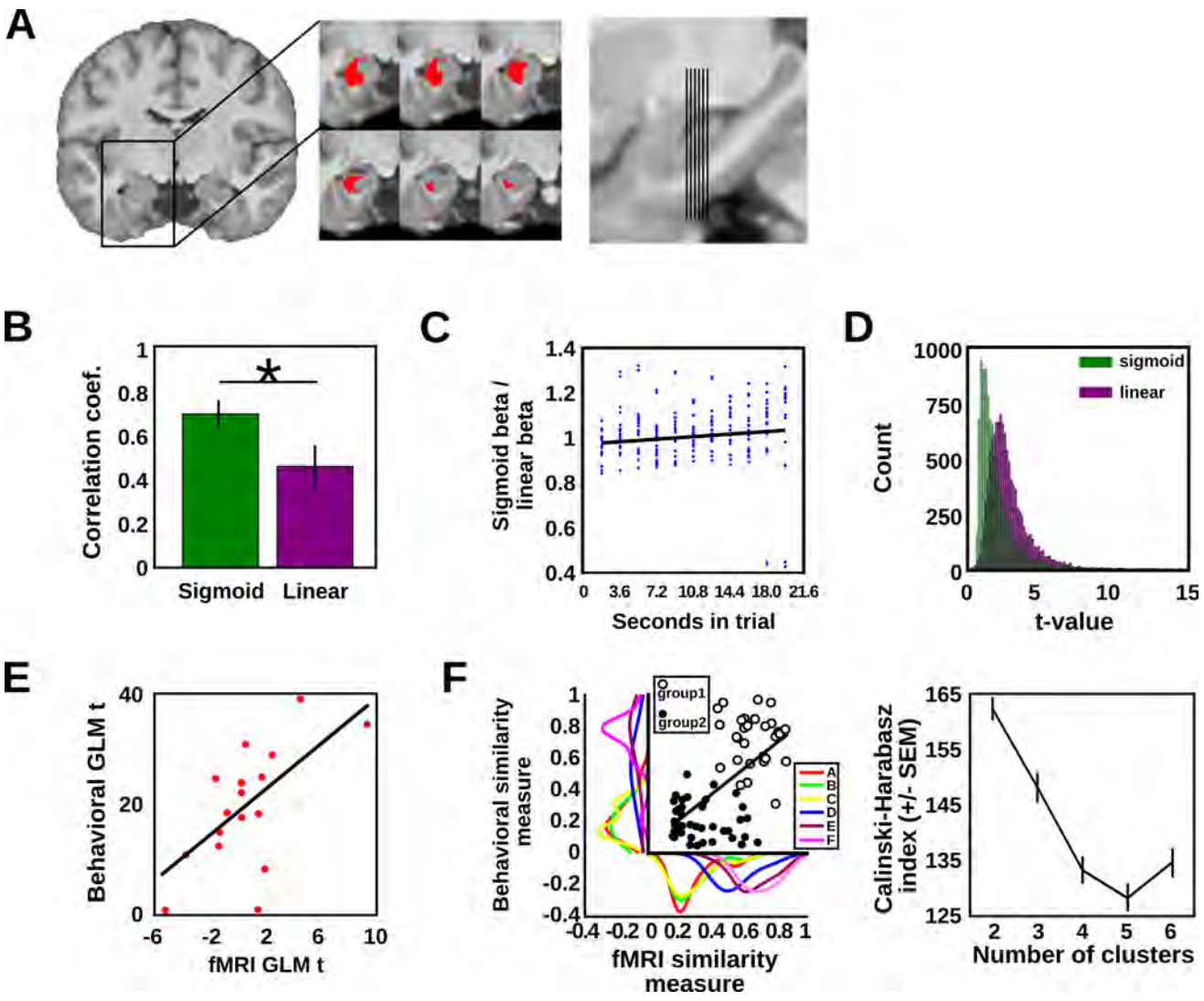

Figure 4. Participants with the strongest sigmoidal effect in behavior also show the strongest sigmoidal response pattern in the hippocampus. (**A**) Both behavioral similarity to base environments (ΔA/(ΔA+ΔF)) and multi-voxel fMRI pattern similarity measures were separately tested against a predictor matrix reflecting an canonical sigmoidal model using GLM, and the resulting t-maps were correlated across participants (see Figure S4B). The same analysis was also performed using a canonical linear predictor matrix. Group effects for the sigmoidal model, restricted to hippocampal region-of-interest and thresholded at a bootstrap corrected P<0.05, are overlaid on a study-specific template. Shown is the extent of the hippocampal effect from y=-16 to y=-10 with slice locations

shown on a sagittal plane. These results were not influenced by differences in navigational behavior or mean hippocampal BOLD signal across environments (see Figure S2). (**B**) Bar plots show the correlation coefficients for the canonical sigmoidal and linear model at their peak voxel in the hippocampus +/- SEM (x=-39, y=-13, z=-21; bootstrap corrected $P<0.05$). (**C**) Multivoxel fMRI data surrounding the hippocampal peak voxel from panel **A** explained by the sigmoidal model relative to the linear model as trials progress (see Supplemental Experimental Procedures). A shift towards a dominantly sigmoidal fit over time was observed, reflected in a significant positive slope of the fit ratio during trial progression (linear regression: slope=0.004, $P<0.05$). The model predicts a fit ratio below 1 up to 5.4 seconds in the trial and above 1 after 9 seconds in the trial. (**D**) Monte-Carlo simulation showing the adherence of the linear and sigmoidal model on randomly shuffled data; higher t-values indicate a better fit. The distribution of t-values from the linear model fits is wider and shifted to higher values compared to the t-values from the sigmoidal model fits (Kolmogorov-Smirnov test: $P<0.001$). (**E**) Scatter plot shows individual t-values from the fMRI against the behavioral GLM using the sigmoidal prediction model for every participant in the hippocampal peak voxel. (**F**) Similarity measures in each environment from both the fMRI and behavioral data are plotted against each other (see Supplemental Experimental Procedures). A significant positive correlation was observed between the two similarity measures (Pearson's correlation: $t_{18}=3.372$, $P<0.001$, $R=0.622$). Environment-wise distribution curves are plotted separately on the x- and y-axis for the fMRI and behavioral similarity measure, respectively. K-means cluster analysis revealed that a two-cluster separation resulted in the highest Calinski-Harabasz index value (right plot). The two clusters in the left plot are indicated by open and closed dots (group 1 and group 2) and are clearly separated along the diagonal. This suggests two basic patterns of data rather than a continuum, indicative of a concurrent sigmoidal pattern in behavioral and neural responses.

References

1. Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. U. S. A. *79*, 2554-2558.

2. Marr, D. (1971). Simple memory: a theory for archicortex. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *262*, 23-81.

3. McClelland, J. L., McNaughton, B. L. and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol. Rev. *102*, 419-457.

4. McNaughton, B. L. and Morris, R. G. M. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system Trends. Cogn. Sci. *10*, 408-415.

5. Norman, K. A. and O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. Psychol. Rev. *110*, 611-646.

6. Doeller, C. F., King, J. A. and Burgess, N. (2008). Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. Proc. Natl. Acad. Sci. U. S. A. *105*, 5915-5920.

7. Ekstrom, A. D., Kahana, M. J., Caplan, J. B., Fields, T. A., Isham, E. A., Newman, E. L. and Fried, I. (2003). Cellular networks underlying human spatial navigation. Nature *425*, 184-188.

8. Howard, L. R., Javadi, A. H., Yu, Y., Mill, R. D., Morrison, L. C., Knight, R., Loftus, M. M., Staskute, L. and Spiers, H. J. (2014). The hippocampus and entorhinal cortex encode the path and Euclidean distances to goals during navigation. Curr. Biol. *24*, 1331-1340.

9. Miller, J. F., Neufang, M., Solway, A., Brandt, A., Trippel, M., Mader, I., Hefft, S., Merkow, M., Polyn, S. M., Jacobs, J. et al. (2013). Neural activity in human hippocampal formation reveals the spatial context of retrieved memories. Science *342*, 1111-1114.

10. Morgan, L. K., Macevoy, S. P., Aguirre, G. K. and Epstein, R. A. (2011). Distances between real-world locations are represented in the human hippocampus. J. Neurosci. *31*, 1238-1245.

11. Spiers, H. J. and Maguire, E. A. (2006). Thoughts, behaviour, and brain dynamics during navigation in the real world. Neuroimage *31*, 1826-1840.

12. Wolbers, T. and Büchel, C. (2005). Dissociable retrosplenial and hippocampal contributions to successful formation of survey representations. J. Neurosci. *25*, 3333-3340.

13. Buckner, R. L. (2010). The role of the hippocampus in prediction and imagination. Annu. Rev. Psychol. *61*, 27-48.

14. Carr, V. A., Rissman, J. and Wagner, A. D. (2010). Imaging the human medial temporal lobe with high-resolution fMRI. Neuron *65*, 298-308.

15. Davachi, L. (2006). Item, context and relational episodic encoding in humans. Curr. Opin. Neurobiol. *16*, 693-700.

16. Eichenbaum, H., Yonelinas, A. P. and Ranganath, C. (2007). The medial temporal lobe and recognition memory. Annu. Rev. Neurosci. *30*, 123-152.

17. Hasselmo M. E. (2012). How we remember: Brain Mechanisms of Episodic Memory (MIT Press, Cambridge)

18. Shohamy, D. and Turk-Browne, N. B. (2013). Mechanisms for widespread hippocampal involvement in cognition. J. Exp. Psychol. Gen. *142*, 1159-1170.

19. Bostock, E., Muller, R. U. and Kubie, J. L. (1991). Experience-dependent modifications of hippocampal place cell firing. Hippocampus *1*, 193-205.

20. Jezek, K., Henriksen, E. J., Treves, A., Moser, E. I. and Moser, M.-B. (2011). Theta-paced flickering between place-cell maps in the hippocampus. Nature *478*, 246-249.

21. Knierim, J. J. and Zhang, K. (2012). Attractor dynamics of spatially correlated neural activity in the limbic system. Annu. Rev. Neurosci. *35*, 267-285.

22. Wills, T. J., Lever, C., Cacucci, F., Burgess, N. and O'Keefe, J. (2005). Attractor dynamics in the hippocampal representation of the local environment. Science *308*, 873-876.

23. Blumenfeld, B., Preminger, S., Sagi, D. and Tsodyks, M. (2006). Dynamics of memory representations in networks with novelty-facilitated synaptic plasticity. Neuron *52*, 383-394.

24. Quian Quiroga, R., Kraskov, A., Mormann, F., Fried, I. and Koch, C. (2014). Single-cell responses to face adaptation in the human medial temporal lobe. Neuron *84*, 363-369.

25. Rotshtein, P., Henson, R. N. A., Treves, A., Driver, J. and Dolan, R. J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. Nat. Neurosci. *8*, 107-113.

26. Doeller, C. F., Barry, C. and Burgess, N. (2010). Evidence for grid cells in a human memory network. Nature *463*, 657-661.

27. Marchette, S. A., Vass, L. K., Ryan, J. and Epstein, R. A. (2014). Anchoring the neural compass: coding of local spatial reference frames in human medial parietal lobe. Nat Neurosci *17*, 1598-1606.

28. Wolbers, T., Wiener, J. M., Mallot, H. A. and Büchel, C. (2007). Differential recruitment of the hippocampus, medial prefrontal cortex, and the human motion complex during path integration in humans. J. Neurosci. *27*, 9408-9416.

29. Norman, K. A., Polyn, S. M., Detre, G. J. and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends. Cogn. Sci. *10*, 424-430.

30. Eichenbaum, H. and Cohen, N. J. (2014). Can we reconcile the declarative memory and spatial navigation views on hippocampal function? Neuron *83*, 764-770.

31. Duncan, K., Sadanand, A. and Davachi, L. (2012). Memory's penumbra: episodic memory decisions induce lingering mnemonic biases. Science *337*, 485-487.

32. Guitart-Masip, M., Barnes, G. R., Horner, A., Bauer, M., Dolan, R. J. and Duzel, E. (2013). Synchronization of medial temporal lobe and prefrontal rhythms in human decision making. J. Neurosci. *33*, 442-451.

33. Kumaran, D., Summerfield, J. J., Hassabis, D. and Maguire, E. A. (2009). Tracking the emergence of conceptual knowledge during human decision making. Neuron *63*, 889-901.

34. Wimmer, G. E. and Shohamy, D. (2012). Preference by association: how memory mechanisms in the hippocampus bias decisions. Science *338*, 270-273.

35. Zeithamova, D., Dominick, A. L. and Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. Neuron *75*, 168-179.

36. Leutgeb, J. K., Leutgeb, S., Treves, A., Meyer, R., Barnes, C. A., McNaughton, B. L., Moser, M.-B. and Moser, E. I. (2005). Progressive transformation of hippocampal neuronal representations in "morphed" environments. Neuron *48*, 345-358.

37. Redish, A. D. (2001). The hippocampal debate: are we asking the right questions? Behav Brain Res *127*, 81-98.

38. Bonnici, H. M., Chadwick, M. J., Kumaran, D., Hassabis, D., Weiskopf, N. and Maguire, E. A. (2012a). Multi-voxel pattern analysis in human hippocampal subfields. Front. Hum. Neurosci. *6*, 290.

39. Bonnici, H. M., Kumaran, D., Chadwick, M. J., Weiskopf, N., Hassabis, D. and Maguire, E. A. (2012b). Decoding representations of scenes in the medial temporal lobes. Hippocampus *22*, 1143-1153.

40. Stokes, J., Kyle, C. and Ekstrom, A. D. (2015). Complementary roles of human hippocampal subfields in differentiation and integration of spatial context. J. Cogn. Neurosci. *27*, 546-559.

41. Lever, C., Wills, T., Cacucci, F., Burgess, N. and O'Keefe, J. (2002). Long-term plasticity in hippocampal place-cell representation of environmental geometry. Nature *416*, 90-94.

42. Agarwal, G., Stevenson, I. H., Berényi, A., Mizuseki, K., Buzsáki, G. and Sommer, F. T. (2014). Spatially distributed local fields in the hippocampus encode rat position. Science *344*, 626-630.

43. Bakker, A., Kirwan, C. B., Miller, M. and Stark, C. E. L. (2008). Pattern separation in the human hippocampal CA3 and dentate gyrus. Science *319*, 1640-1642.

44. Kyle, C. T., Stokes, J. D., Lieberman, J. S., Hassan, A. S. and Ekstrom, A. D. (2015). Successful retrieval of competing spatial environments in humans involves hippocampal pattern separation mechanisms. Elife *4*,

45. Horner, A. J., Bisby, J. A., Bush, D., Lin, W.-J. and Burgess, N. (2015). Evidence for holistic episodic recollection via hippocampal pattern completion. Nat. Commun. *6*, 7462.

Supplemental Figures

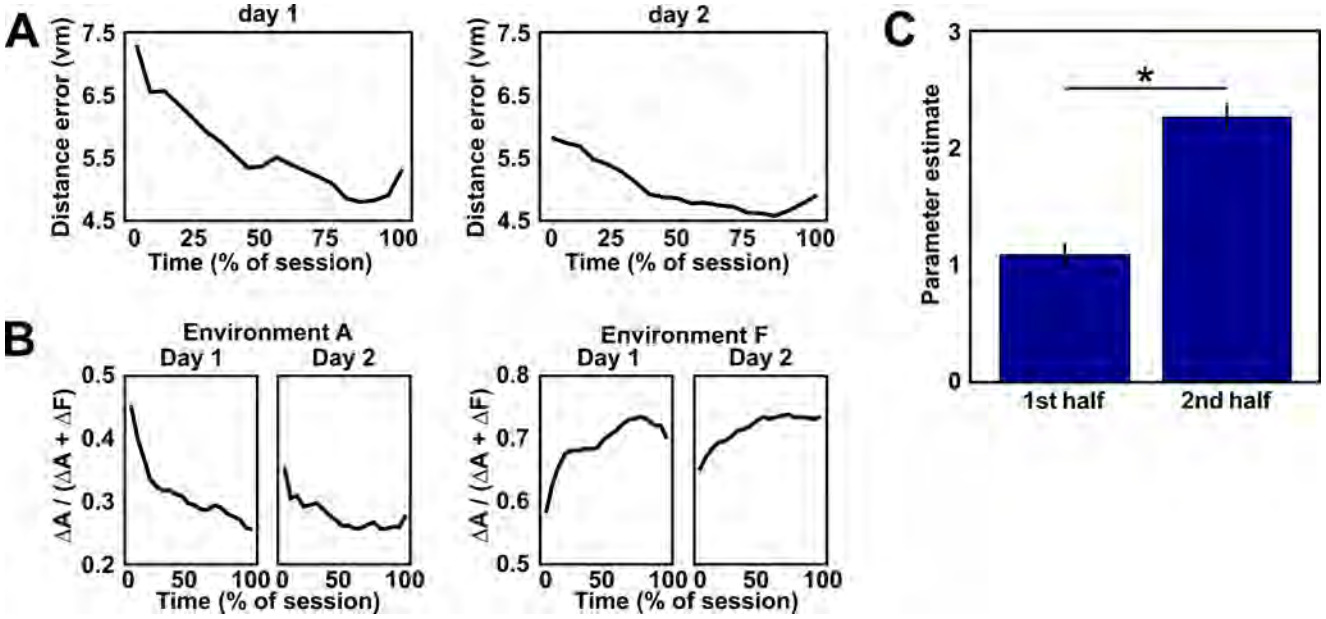

Figure S1, related to Figure 1, learning phase. (**A**) Mean distance error (distance between response location and object location measured in virtual meters (vm) and averaged over objects and participants) decreased over time in both environment A and F (mean decrease in drop error on day 1 = 0.020 vm/trial, on day 2 = 0.008 vm/trial; average moment of asymptote in % session duration on day 1 = 67.1%, on day 2 = 56.8%). On the first day, participants performance improved more in the first half of the session than the second half, as defined by the slope of learning (paired t-test t_{19}=2.18, P<0.05) but not on the second day (paired t-test t_{19}=1.30, P=0.21). (**B)** Plots show the time course of the behavioral similarity to baseline environments (ΔA/(ΔA+ΔF), see Supplemental Experimental Procedures) reflecting the 'A-ness' (below 0.5) and 'F-ness' (above 0.5) of the memory response. The increase in environment-specific memory response is most evident during the first day of training. An asymptote of the learning rate, reflecting maximum performance on a given day, was observed on both days in each environment for 10 participants. Six participants only showed an asymptote for both environments on day 2, suggesting they did not reach maximum performance on day 1. When averaging the locations of the asymptotes from day 2 over the two environments, participants reached asymptotic learning after 55.6% of the total session duration. Three participants showed only an asymptote on day 1, further inspection revealed that these participants had stable performance during the complete session on day 2. One participant had no asymptote and did not improve on either day. Paired t-test on the averaged performances of the last 50% of trials on the second training day revealed no difference in performance between environments A and F (t_{19}=1.19, P=0.25). (**C**) fMRI data (n=11) from an exploratory analysis in the hippocampus obtained during the first training days was temporally split (2 x 30 minutes) and multi-voxel correlations between environments A and F during navigation were tested against the model that A differs from F. Bar plots show resulting parameter estimates (averaged across participants +/- SEM), for the first and second half. Parameter estimates differed between the two halves of the session (t_{10} =2.53, P<0.03), indicating that the correlation of multi-voxel patterns between A and F decreased during training.

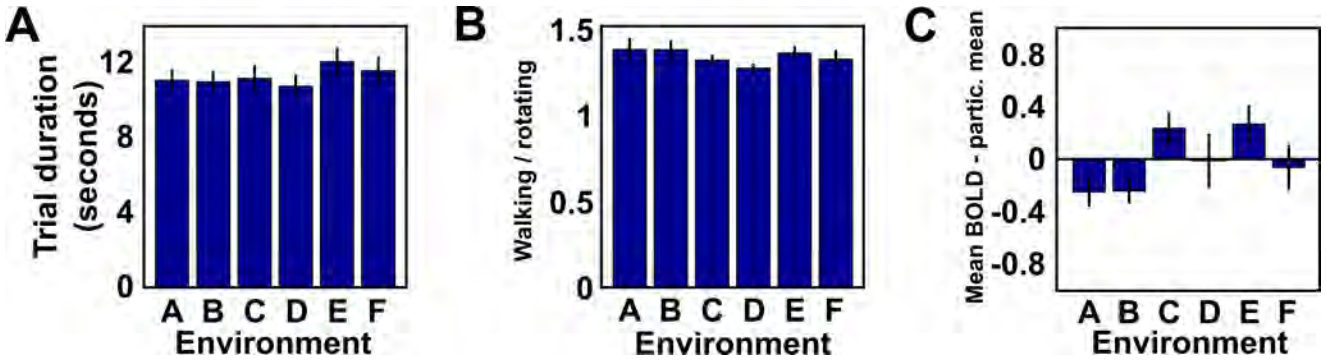

Figure S2, related to Figure 3 and 4: Neither navigation behavior nor mean BOLD signal differs between environments. Bars show (**A**) mean time spent in each environment in seconds, (**B**) time spent walking relative to time spent rotating and (**C**) mean BOLD signal for the six environments (all bars averaged across participants +/- SEM). For each participant, the mean BOLD signal across all environments was subtracted from the environment-specific BOLD signal to account for the differences in baseline activity between participants. No differences between environments were observed (ANOVA **A**: $F_{(5,95)}$=1.88, P=0.11; **B**: $F_{(5,95)}$=1.17, P=0.33; **C**: $F_{(5,95)}$=0.12, P=0.98).

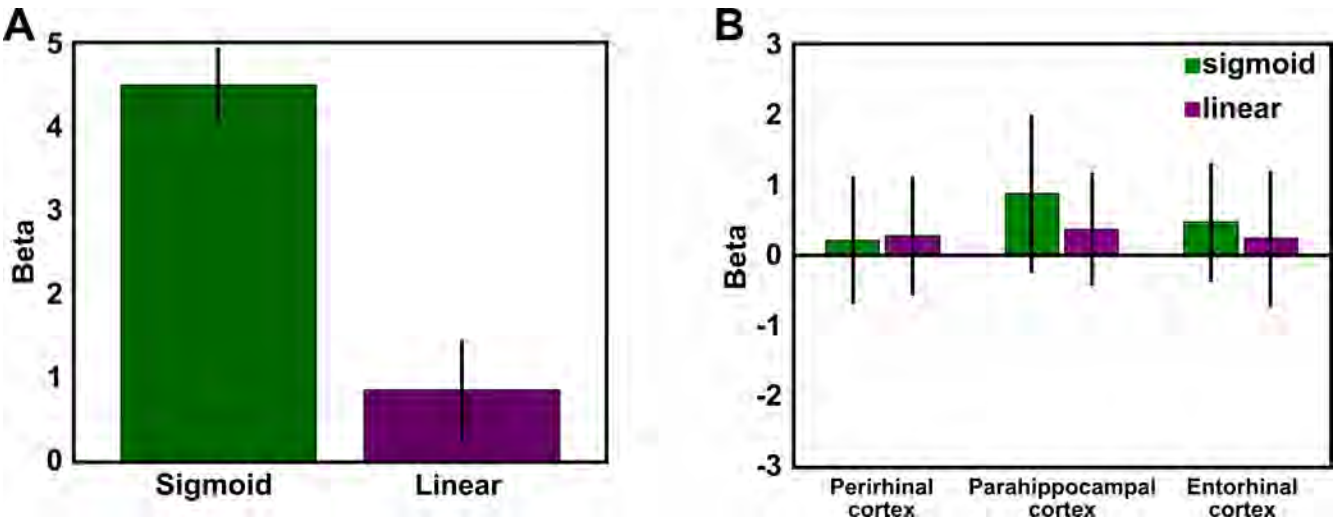

Figure S3, related to Figure 3: Sigmoidal fMRI effect is not driven by path similarity and is neither present in extrahippocampal cortices within the medial temporal lobe. (**A**) To test whether the observed increased multi-voxel correlation between environments A, B, and C and between environments D, E, and F is partially driven by similarity in the paths walked or object properties, we excluded same-object comparisons from the RSA model. This means that in any 2 trials where the object to be placed was the same, no prediction was made on the similarity of the fMRI data between these trials. We compared this model to the fMRI data at the peak voxel from our behavioral informed RSA analysis (see Figure 3), bars show the resulting beta estimates +/- SEM. The analysis revealed a significant difference remained between the sigmoidal and linear models (paired t-test on the RSS from the 2 model fits: t_{19}=2.81, P=0.006). (**B**) Both a perfect sigmoidal and linear model were tested on three extrahippocampal regions of interest within the medial temporal lobe, the bars show the resulting beta estimates +/- SEM. For all ROIs, for all models, beta estimates did not significantly differ from zero (all P>0.36).

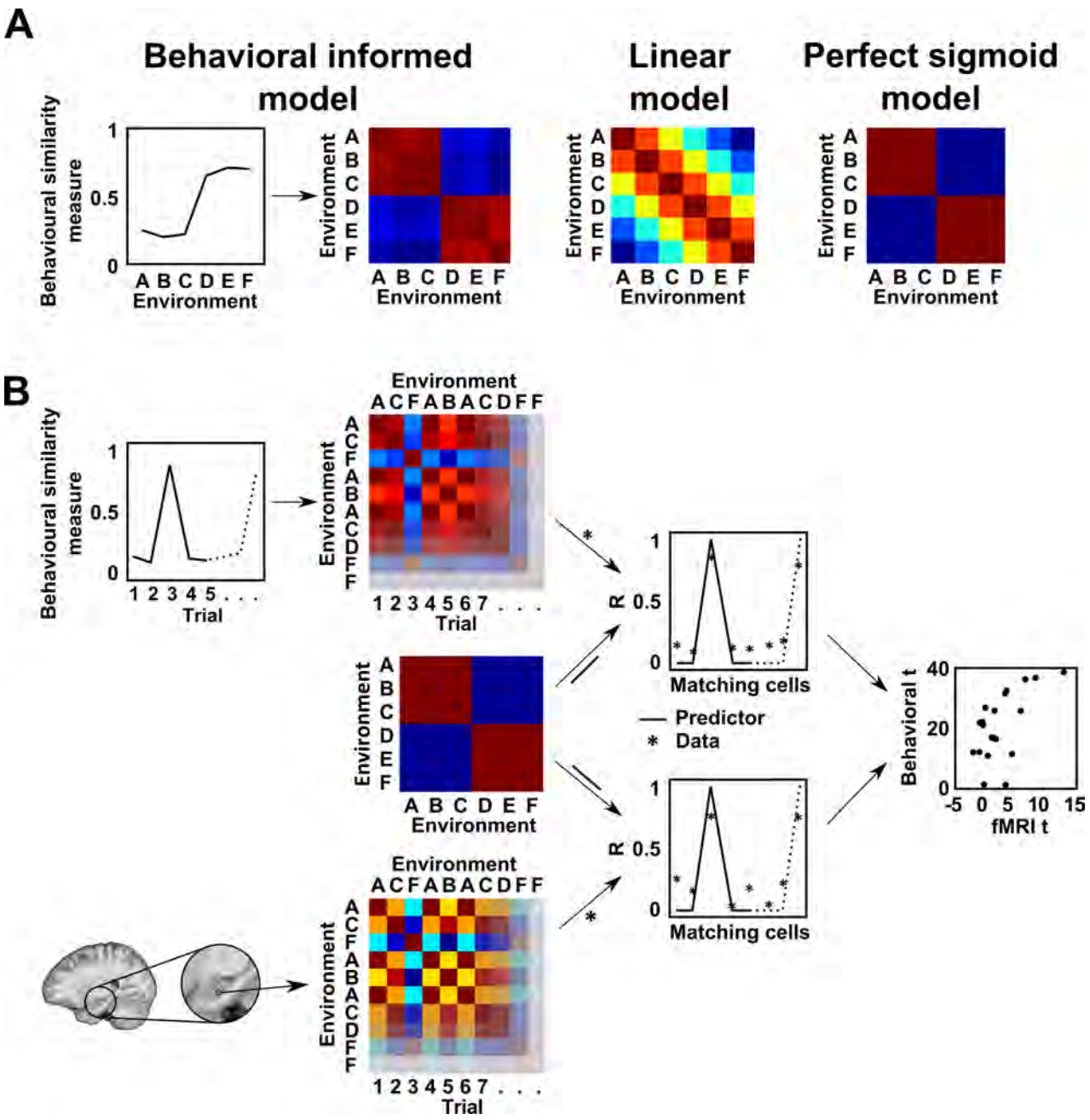

Figure S4, related to Figures 3 and 4. Schematic of analysis logic testing the correlation between behavioral and fMRI sigmoidal response profiles. (**A**) Left: Example of predictor matrices used in behavioral informed analysis. Middle: linear predictor matrix which did not change between analyses. Right: Sigmoidal model predicting perfect similarity between A, B and C, and separately between D, E and F, while also predicting perfect separability between A, B, C and D, E, F. (**B**) Behavioral similarity to baseline environments (example in top left panel) and multi-voxel patterns in searchlights (bottom left) were used to generate separate trial-by-trial correlation matrices. These two data matrices were separately tested against a predictor matrix reflecting a perfect sigmoidal model using GLM. The two resulting arrays of t-values were correlated across participants (right: example data for illustrative purpose). The same analysis was also performed using a perfect linear predictor matrix (see Figure 4).

Supplemental Experimental Procedures

Participants. 25 participants took part in this study of which 3 participants were excluded from further data analyses due to excessive head motion during scanning (high occurrence of >5mm movement and >3 degrees rotation spikes during a session). In addition, 2 participants were excluded due to scanner malfunctioning. Of the remaining 20 participants, 6 were male and 14 female, with an average age of 21.3 years (ranging between 19 and 28). Participants gave written consent and were paid for participating, as approved by the local Research Ethics Committee (CMO region Arnhem-Nijmegen, The Netherlands). All participants had normal or corrected-to-normal vision and reported to be in good health.

Virtual reality environments. We developed a virtual reality (VR) task using the UnrealEngine2 Runtime software (http:/udn.epicgames.com/Two/Webhome.html). Within this software we created a virtual arena that comprised a circular grassy plane surrounded by a wall with mountains in the background. By systematically changing the shape of the mountains we created six different environments (A through F). Since hippocampal fMRI responses are sensitive to changes in environmental geometry [S1-2], we shaped the mountains such that they linearly morphed from A through F (B = 20%A + 80%F; C = 40%A + 60%F; D = 60%A + 40%F; E = 80%A + 20%F.; see Figure 1B). The backgrounds of the two extreme environments (A and F) were chosen such that they are clearly discriminable but differ in as few features as possible to allow for smooth morphing between them. This resulted in both these backgrounds containing two prominent mountains of different heights and the same feature-restrained sky. Background images were created by generating height maps using Matlab 7.9 (http://www.mathworks.com) which were transformed into images of landscapes using the Terragen software package (http://www.planetside.co.uk). In the VR environment participants embodied an avatar with first person perspective and moved by using a button-box with their right hand enabling forward movement and left and right turns. The avatar's location and heading were recorded every 100 milliseconds.

Virtual reality task. Our task consisted of a two day learning phase (separated by 23 hours) immediately followed by a testing phase. In the learning phase participants learned the location of four objects in two separate environments, each object having a unique locations in each environment (Figure 1A). In the testing phase participants replaced the objects in both the learned environments (A and F) and the four novel morph environments (B, C, D and E).

The learning phase started with a short initial familiarization phase in which the four objects were shown once in environment A and once in F. In each subsequent trial an object was cued by appearing in the top part of the screen for 2 seconds after which participants navigated to the location they thought the object should be (time-out after 20 seconds, mean trial duration = 14.63 seconds). Feedback was provided by showing the object at its correct position after which participants had to collect the object (mean duration = 6.42 seconds; see Figure 1C). Every 9 minutes environments were switched by spawning a navigable bridge connecting the two environments that the participants had to cross. Participants visited each environment 3 times on each training day (first day: mean number of trials = 159.0, SEM = 6.8. second day: mean number of trials = 149.7, SEM = 8.6). In the testing phase participants performed between 1 and 4 trials before an ITI of 2-4 seconds appeared. Unknown to the participants, environments were changed during an ITI to one of the six environments in the morph sequence. Trial structure for the testing phase was unchanged other than that no feedback on the true object location was provided (mean number of trials = 235.5, SEM = 7.7, mean trial duration = 12.38 seconds). The order of environments and objects was counterbalanced so that no object would be cued twice in a row and the unique object-environment combinations are distributed uniformly over a session. In addition, for 15 participants we recorded a confidence rating (on a 5 point scale) on the precision of placement of the object after each trial.

Analysis of behavioral data. Spatial memory performance during all sessions was measured as the Euclidean distance between the response location and the correct object location in each trial. In order to test if participants' spatial memories were stable after the learning session, we fitted this 'drop error' over time of both training sessions to a quadratic polynomial model. The differential of the resulting function determined the slope at any given point on the learning curve. An average negative slope indicated learning of the location of the objects while the existence of a horizontal asymptote indicated maximum acquisition on a given day (Figure S1).

During the testing phase, our main behavioral measure was the relative difference between the object replacement location and the true object location in environments A (ΔA) and F (ΔF), calculated as $\Delta A/(\Delta A+\Delta F)$. This behavioral similarity measure scales linearly from 0 to 1 with increasing ΔA and decreasing ΔF (Figure 1D), reflecting the 'A-ness' and 'F-ness' of each memory response. This measurement was fitted to a sigmoid and a linear model using Maximum Likelihood

Estimation (MLE), for which the probability density functions were defined as six normal distributions with the mean of each distribution corresponding to values predicted by variable parameters for a linear or a sigmoid model. The standard deviation of all normal distributions was fixed to 0.25 so that at least 95% of the possible values would fall within the probability distribution. The resulting probability values were normalized so the area under the curve of the probability distributions summed to 1. Both the linear and the sigmoidal model have 2 free parameters, thus model complexity was held constant. The sigmoidal model was defined as

$$y_i = \frac{1+a}{1+e^{-x_i+b}+(1+a)/2}$$

where a is the amplitude and b the horizontal offset. The linear model was defined as

$$y_i = a * x_i + b$$

where a is the slope and b the vertical offset. Per participant, we fitted each model to the behavioral data and compared the resulting residual sum of square (RSS) between the models using a paired t-test.

Behavioral control experiment - Perception of background cues. We performed a separate behavioral experiment to test if there was non-linearity in the perception of the environments' background images (from A to F). 16 naive participants (12 females and 4 males, average age: 22.0 years; range: 16 to 29 years) were presented with 392 trials each consisting of image pairs representing the background of two environments. The images were displayed consecutively (stimulus duration lasted 2 seconds for each image) with a 1 second mask image in between (scrambled version of the first image of each pair). At the end of each trial, participants scored the perceived difference between the presented images on a 5-point similarity scale, followed by an ITI of 1 second, see Figure 2A. Trials were divided into 8 blocks separated with 20 seconds breaks. We applied to the similarity scores the same analysis logic as to the behavioral memory responses from the main experiment, i.e. the scored difference of each environment compared to either environment A (ΔA) or environment F (ΔF) represented as ΔA/(ΔA+ΔF). This measure was fitted to a sigmoid and linear model using MLE (see above). These models were then compared using a paired t-test on the resulting RSS.

MRI data acquisition. Imaging data was acquired on a Siemens 3T Trio scanner using a 32- channel head coil. The functional sequence used was a custom multi-echo 3D EPI sequence (TR = 1800 ms; TE = 25 ms; flip angle = 15°; 64 slices of matrix 112 × 112 with a 25% gap; voxel size = 2 × 2 × 2 mm). Scanning of functional images was subdivided into two blocks of 30 minutes each (1000 volumes each) with a break of approximately five minutes. In addition to the testing session, 11 out of the 20 participants were scanned during the first training session. For every scanning session, a field map using a gradient echo sequence was recorded for distortion correction of the acquired EPI images. The structural scan comprised a MPRAGE-grappa sequence (TR = 2300 ms; TE = 3.03 ms; flip angle = 8°; in-plane resolution = 256x256 mm; number of slices = 192; acceleration factor PE = 2; voxel resolution = 1 mm^3, duration = 321 seconds).

Image preprocessing. For the fMRI analysis, we used the Automatic Analysis framework [S3], which combines tools from SMP8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/), FreeSurfer (http://surfer.nmr.mgh.harvard.edu/) and the FMRIB Software Library v5.0 (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/), complemented by custom scripts executed in Matlab 7.9. Preprocessing consisted of de-noising of the structural image using a non-local means filter [S4], motion correction, co-registration of functional images to the structural scan and brain-extraction. White matter was masked out using FreeSurfer. Finally, a study specific structural template was build and non-linear normalization of participant-specific contrasts to this template was performed using the Advanced Normalization Tools version 1.9 for Unix (http://www.picsl.upenn.edu/ANTS).

Representational similarity analysis. fMRI data were modeled with square wave regressors, one regressor for each trial. The start of each trial was defined by the onset of a cue indicating which object should be relocated by the participant and the response by the participant defined the end of each trial. The average trial duration was 14.63 seconds (i.e. 8.1 TRs). The regressors were convolved with a hemodynamic response function (HRF) and voxel-wise unsmoothed beta values were extracted per regressor of interest. Additional nuisance regressors consisted of three translations and three rotations derived

from motion correction, as well as single spike regressors reflecting large, sudden, head movements. The voxel-wise betas were used in subsequent searchlight analyses using representational similarity analysis (RSA) [S5]. RSA comprised the voxel-wise extraction of beta values from each trial, which were correlated across trials resulting in a trial by trial correlation matrix. For every combination of trials a prediction on their similarity was made based on the environments navigated in those trials (Figure 3A). How well the similarity matrices fit the prediction models was assessed using general linear modeling (GLM). The resulting searchlight voxel-wise beta brain maps were normalized, smoothed and subjected to second-level random effects analysis.

Group-level fMRI analysis. In a first analysis on fMRI data from the testing phase (see Figure 3), the sigmoid fitted to each participant's behavioral similarity measures $\Delta A/(\Delta A+\Delta F)$ was used to predict the similarity in the multi-voxel fMRI pattern on a trial-by-trial basis. The resulting predictor matrix (see Figure S4B) was subsequently tested against the actual similarity between different trials from the fMRI data. A linear model was used as control since this adhered to the visual changes in the environment morph sequence (Figure 2C; see Behavioral control experiment above).

Given our strong a prior hypothesis, analyses were initially restricted to the hippocampus, details on the region-of-interest are outlined below. Using a bootstrap method we applied a corrected statistical threshold of $P<0.05$ to all GLM results. A bootstrap method was also used to directly compare the fit between these models while keeping multiple comparison correction (see Bootstrap analysis below). To assess the relationship between putative attractor dynamics in both behavior and neural pattern (see Figure 4), we compared how well a perfect sigmoidal model (step-function) fit the behavioral and fMRI data, using again a linear model as control (Figure S4A). Fitting of models to both the behavioral data ($\Delta A/(\Delta A+\Delta F)$ and fMRI data (multi-voxel trial- by-trial fMRI pattern) was performed using GLM. Subsequently, resulting t-maps were correlated over participants, separately for sigmoidal and linear models (Figure 4B). Again, a bootstrap method was used to assess the significance of the correlations while keeping multiple comparison correction (see Bootstrap analysis below).

Bootstrap analysis. To assess whether fMRI data in any voxel significantly fit a sigmoidal or linear model we shuffled all the first level contrast images and applied a 5 mm half width half maximum variance smoothing to the contrasts. We did this 5000 times and for each permutation the resulting data was tested against the sigmoidal and linear prediction model generated as if the data was non-shuffled. The resulting distribution of t-values were used for testing the significance of fit of either model to non-shuffled data. A similar bootstrap method was used to directly compare the fit between these models while keeping multiple comparison correction. Per participant we subtracted the GLM t-value of the sigmoidal model from the t-values of the linear model and using a one-sample t-test on this difference a p-value was calculated. This step was repeated for 5000 permutations, each time shuffling the voxels, and the resulting distribution of p-values provided the critical value to test the p-value of a one-sample t-test on non-shuffled data. A bootstrap method was also used to assess where in the hippocampus significant correlation between fMRI data and behavioral data existed. This method comprised shuffling the t-values obtained by applying a GLM on fMRI data using either a perfect sigmoidal or linear model over participants for every voxel. The resulting t-values were correlated with the t-values resulting from fitting the same models to non-shuffled behavioral data. This process was repeated 5000 times and rendered a distribution of correlation coefficients per voxel used to test the significance of the correlation coefficient between the non-shuffled fMRI and behavioral data.

Regions Of Interest. A hippocampal ROI was manually segmented using MeVisLab software (MeVis Medical Solutions AG, Bremen, Germany) on a study specific structural template generated by the ANTS software (see above). The segmentation was based on a protocol described by [S6] and [S7], and guided by an anatomical atlas of the human hippocampus [S8]. Post-hoc ROI analyses using the perfect sigmoidal and linear model were performed on the perirhinal, entorhinal and parahippocampal cortex. For the perirhinal cortex ROI, a probabilistic map was used [S9]. The entorhinal cortex ROI [S10] was obtained from the SPM anatomy toolbox version 2.2 and the parahippocampal cortex ROI was obtained from the Talairach label database [S11-12] and transformed into MNI space using BioImage Suite version 3.01 (http://bioimagesuite.yale.edu). All probabilistic maps were visually inspected to ensure that no hippocampal voxels were included and a relative conservative threshold was applied (voxels included were in >90% of brains used to make the probabilistic maps labeled as the ROI).

Searchlight analysis. Searchlight mapping was performed on the native space images of each participant by moving a spherical searchlight (4 voxel radius) through the gray-matter masked volume or ROI one voxel at a time. Statistics were mapped back to the central voxel of each spherical searchlight thus yielding a single-participant information map. Analysis

was restricted to searchlights that contained at least 30 voxels, thereby eliminating searchlights that were close to the edge of gray matter. The first-level results were normalized and smoothed with a 4mm FWHM kernel, and a second-level model on the resulting data was carried out to examine information at the group level. All results are reported in MNI coordinates.

Divergence of the neural pattern between environments during training. In a post-hoc analysis on the training phase data, we tested if the neural pattern in the hippocampus changed as a function of learning object positions while navigating environments A and F. We split the data into two halves (first and last 30 minutes of the session) and calculated per trial the voxel-wise correlations with every other trial within each half. The resulting correlations were related to the appropriate cell of an environment-by-environment correlation matrix and this matrix was tested against a matrix predicting a larger neural similarity for activation patterns within an environment than between environments (Figure S1). We predicted that the model would fit the second half of the learning phase better than the first half as the representation of the environments was thought to diverge over time, consistent with attractor dynamics.

Monte-Carlo simulation. In order to test if a sigmoidal model fits the fMRI data better regardless of the specific environment navigated or the trial number, we performed a Monte-Carlo simulation. From the whole data set each voxel from each EPI image was assigned a unique number. From this pool 105750 voxels were randomly picked with replacement and assigned a trial number between 1 and 235 and an environment number between 1 and 6 such that each combination of trial and environment had the same amount of voxels. Using this data, we fitted both a sigmoidal and a linear model as described above and saved the resulting t-values. The process of picking voxels, assigning trials and scenes and fitting both models was repeated 10,000 times resulting in a distribution of t-values for each model, which were compared using the non-parametric Kolmogorov-Smirnov test.

fMRI and behavioral data similarity. To examine the relation between behavioral and fMRI data we calculated a similarity measure for the fMRI data akin to the similarity measure used to analyze the behavioral data. Using GLM, for each environment we modeled both similarity to environment A (ΔA) and similarity to environment F (ΔF). Using the peak voxel from our correlation analysis (see Figure 4A) as our coordinate, we applied the formula $\Delta A/(\Delta A+\Delta F)$ which resulted in a similarity measure that scales linearly from 0 to 1 with increasing ΔA and decreasing ΔF. Having both a similarity measure from behavioral and fMRI data that can be readily compared, we plotted these measures against each other and applied K-means clustering on the resulting 2D space, assessing the optimal number of clusters using the Calinski-Harabasz index value [S13], and examined the correlation between the measures using Pearson's correlation coefficient. K-means clustering was performed 1000 times, each time with new random starting centroids to avoid convergence to a local minima and the result with the lowest total within-cluster point-to-centroid distances is shown in Figure 4C.

Within-trial dynamics of sigmoidacity versus linearity. In a post-hoc analysis we looked at within-trial dynamical changes of the explanatory power of both a perfect sigmoidal and linear model over time, again around the peak voxel obtained from our correlation analysis (see Figure 4A). To this end, we split each trial in overlapping 3.6 second segments with the center of the segments 1.8 seconds apart from neighboring segments. Using GLM, all segments that were in the same time-window within all trials were fitted to a perfect sigmoidal model and the resulting betas were divided by betas obtained from fitting the same data to a linear model. This resulted in a measure directly comparing the contribution of the sigmoidal model to the linear model as trials progressed. This was done for each participant, resulting in no more than 20 data-points per time-window. Using a linear regression we assessed whether this measure changed structurally within trials as they progressed.

Supplemental References

S1. Bird, C. M., Capponi, C., King, J. A., Doeller, C. F. and Burgess, N. (2010). Establishing the boundaries: the hippocampal contribution to imagining scenes. J. Neurosci. *30*, 11688-11695.

S2. Stokes, J., Kyle, C. and Ekstrom, A. D. (2015). Complementary roles of human hippocampal subfields in differentiation and integration of spatial context. J. Cogn. Neurosci. *27*, 546-559.

S3. Cusack, R., Vicente-Grabovetsky, A., Mitchell, D. J., Wild, C. J., Auer, T., Linke, A. C. and Peelle, J. E. (2014). Automatic analysis (aa): efficient neuroimaging workflows and parallel processing using Matlab and XML. Front.

Neuroinform. *8*, 90.

S4. Coupe, P., Yger, P., Prima, S., Hellier, P., Kervrann, C. and Barillot, C. (2008). An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. IEEE. Trans. Med. Imaging. *27*, 425-441.

S5. Kriegeskorte, N., Mur, M. and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4.

S6. Malykhin, N. V., Lebel, R. M., Coupland, N. J., Wilman, A. H. and Carter, R. (2010). In vivo quantification of hippocampal subfields using 4.7 T fast spin echo imaging. Neuroimage *49*, 1224-1230.

S7. Wisse, L. E. M., Gerritsen, L., Zwanenburg, J. J. M., Kuijf, H. J., Luijten, P. R., Biessels, G. J. and Geerlings, M. I. (2012). Subfields of the hippocampal formation at 7 T MRI: in vivo volumetric assessment. Neuroimage *61*, 1043-1049.

S8. Duvernoy H. M., Cattin F. and Risold P. Y. (2013). The Human Hippocampus: Functional Anatomy, Vascularization and Serial Sections with MRI (Berlin: Springer)

S9. Holdstock, J. S., Hocking, J., Notley, P., Devlin, J. T. and Price, C. J. (2009). Integrating visual and tactile information in the perirhinal cortex. Cereb Cortex *19*, 2993-3000.

S10. Amunts, K., Kedo, O., Kindler, M., Pieperhoff, P., Mohlberg, H., Shah, N. J., Habel, U., Schneider, F. and Zilles, K. (2005). Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: intersubject variability and probability maps. Anat Embryol (Berl) *210*, 343-352.

S11. Lancaster, J. L., Rainey, L. H., Summerlin, J. L., Freitas, C. S., Fox, P. T., Evans, A. C., Toga, A. W. and Mazziotta, J. C. (1997). Automated labeling of the human brain: a preliminary report on the development and evaluation of a forward-transform method. Hum Brain Mapp *5*, 238-242.

S12. Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., Kochunov, P. V., Nickerson, D., Mikiten, S. A. and Fox, P. T. (2000). Automated Talairach atlas labels for functional brain mapping. Hum Brain Mapp *10*, 120-131.

S13. Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis Communications in Statistics-theory and Methods *3*, 1-27.

Supplemental Figures

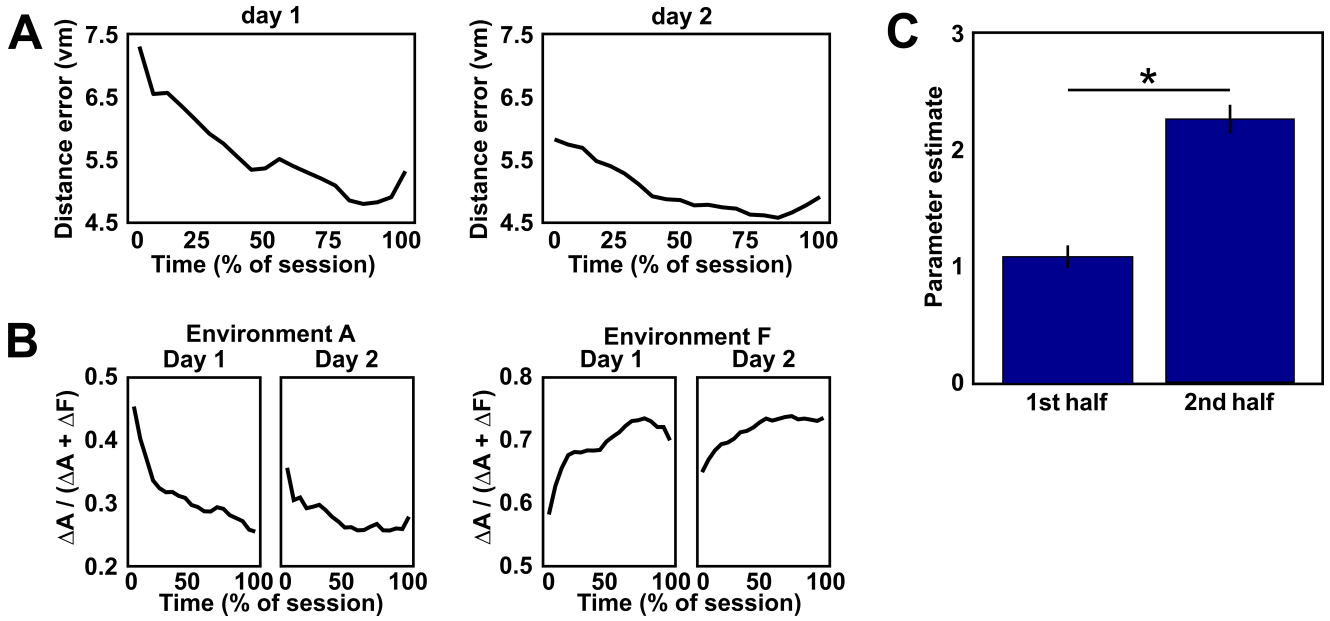

Figure S1, related to Figure 1, learning phase. (**A**) Mean distance error (distance between response location and object location measured in virtual meters (vm) and averaged over objects and participants) decreased over time in both environment A and F (mean decrease in drop error on day 1 = 0.020 vm/trial, on day 2 = 0.008 vm/trial; average moment of asymptote in % session duration on day 1 = 67.1%, on day 2 = 56.8%). On the first day, participants performance improved more in the first half of the session than the second half, as defined by the slope of learning (paired t-test t_{19}=2.18, P<0.05) but not on the second day (paired t-test t_{19}=1.30, P=0.21). (**B)** Plots show the time course of the behavioral similarity to baseline environments (ΔA/(ΔA+ΔF), see Supplemental Experimental Procedures) reflecting the 'A-ness' (below 0.5) and 'F-ness' (above 0.5) of the memory response. The increase in environment-specific memory response is most evident during the first day of training. An asymptote of the learning rate, reflecting maximum performance on a given day, was observed on both days in each environment for 10 participants. Six participants only showed an asymptote for both environments on day 2, suggesting they did not reach maximum performance on day 1. When averaging the locations of the asymptotes from day 2 over the two environments, participants reached asymptotic learning after 55.6% of the total session duration. Three participants showed only an asymptote on day 1, further inspection revealed that these participants had stable performance during the complete session on day 2. One participant had no asymptote and did not improve on either day. Paired t-test on the averaged performances of the last 50% of trials on the second training day revealed no difference in performance between environments A and F (t_{19}=1.19, P=0.25). (**C**) fMRI data (n=11) from an exploratory analysis in the hippocampus obtained during the first training days was temporally split (2 x 30 minutes) and multi-voxel correlations between environments A and F during navigation were tested against the model that A differs from F. Bar plots show resulting parameter estimates (averaged across participants +/- SEM), for the first and second half. Parameter estimates differed between the two halves of the session (t_{10} =2.53, P<0.03), indicating that the correlation of multi-voxel patterns between A and F decreased during training.

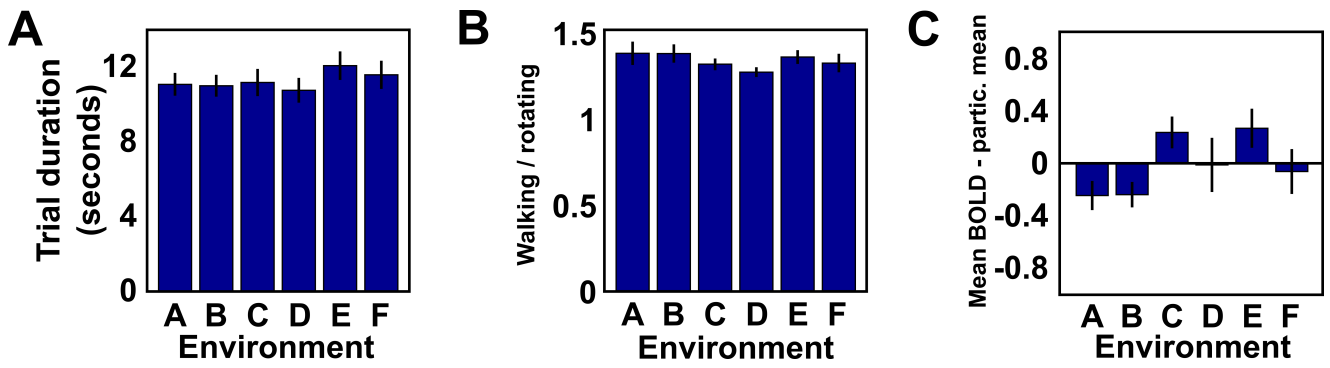

Figure S2, related to Figure 3 and 4: Neither navigation behavior nor mean BOLD signal differs between environments. Bars show (**A**) mean time spent in each environment in seconds, (**B**) time spent walking relative to time spent rotating and (**C**) mean BOLD signal for the six environments (all bars averaged across participants +/- SEM). For each participant, the mean BOLD signal across all environments was subtracted from the environment-specific BOLD signal to account for the differences in baseline activity between participants. No differences between environments were observed (ANOVA **A**: $F_{(5,95)}$=1.88, P=0.11; **B**: $F_{(5,95)}$=1.17, P=0.33; **C**: $F_{(5,95)}$=0.12, P=0.98).

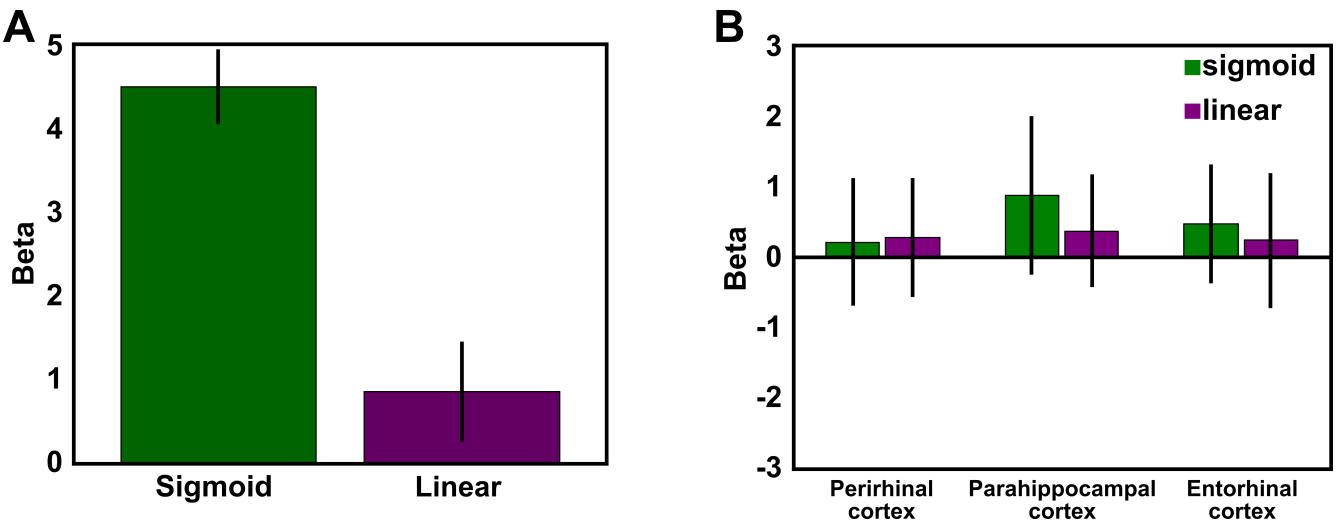

Figure S3, related to Figure 3: Sigmoidal fMRI effect is not driven by path similarity and is neither present in extrahippocampal cortices within the medial temporal lobe. (**A**) To test whether the observed increased multi-voxel correlation between environments A, B, and C and between environments D, E, and F is partially driven by similarity in the paths walked or object properties, we excluded same-object comparisons from the RSA model. This means that in any 2 trials where the object to be placed was the same, no prediction was made on the similarity of the fMRI data between these trials. We compared this model to the fMRI data at the peak voxel from our behavioral informed RSA analysis (see Figure 3), bars show the resulting beta estimates +/- SEM. The analysis revealed a significant difference remained between the sigmoidal and linear models (paired t-test on the RSS from the 2 model fits: t_{19}=2.81, P=0.006). (**B**) Both a perfect sigmoidal and linear model were tested on three extrahippocampal regions of interest within the medial temporal lobe, the bars show the resulting beta estimates +/- SEM. For all ROIs, for all models, beta estimates did not significantly differ from zero (all P>0.36).

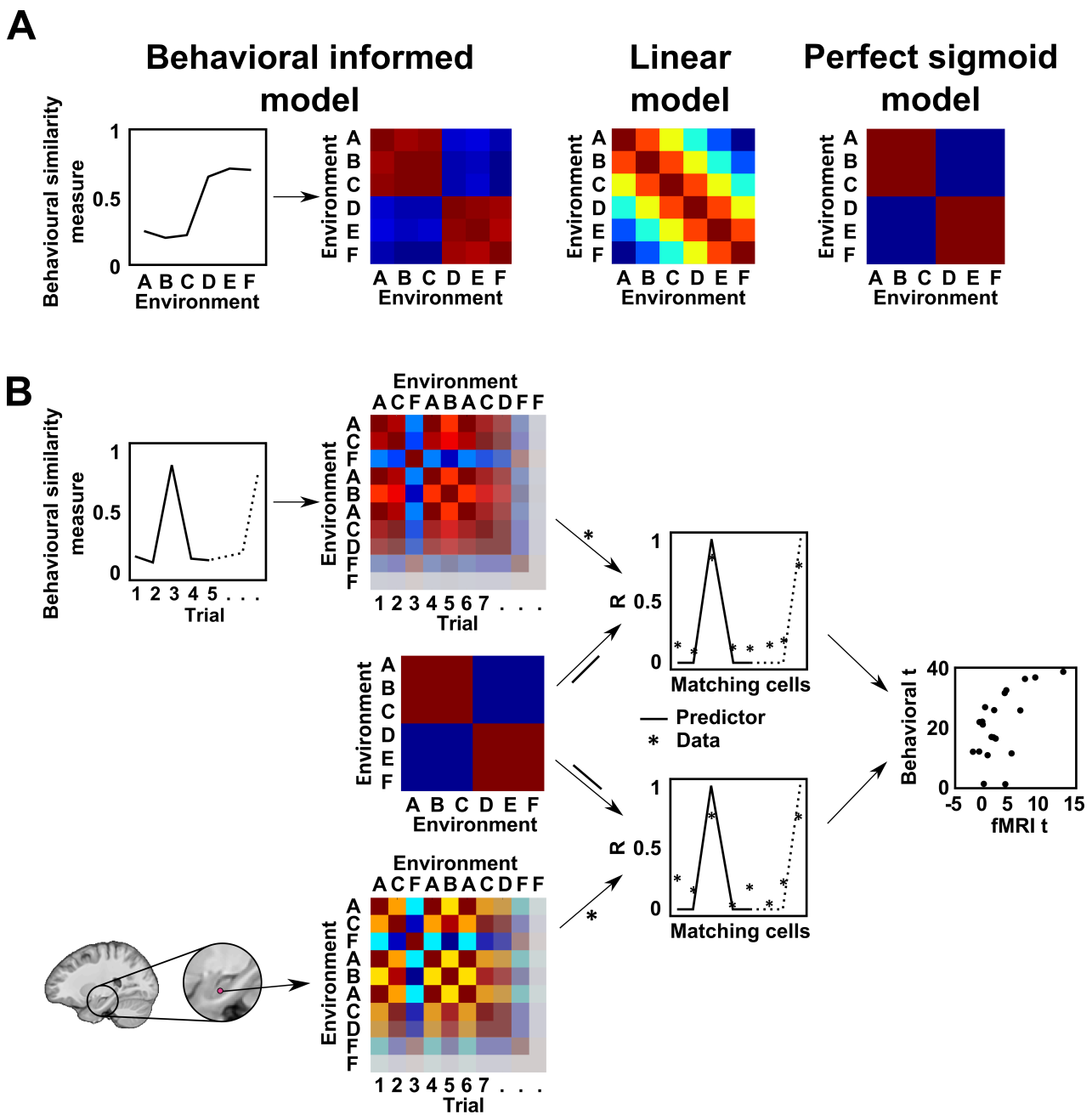

Figure S4, related to Figures 3 and 4. Schematic of analysis logic testing the correlation between behavioral and fMRI sigmoidal response profiles. (**A**) Left: Example of predictor matrices used in behavioral informed analysis. Middle: linear predictor matrix which did not change between analyses. Right: Sigmoidal model predicting perfect similarity between A, B and C, and separately between D, E and F, while also predicting perfect separability between A, B, C and D, E, F. (**B**) Behavioral similarity to baseline environments (example in top left panel) and multi-voxel patterns in searchlights (bottom left) were used to generate separate trial-by-trial correlation matrices. These two data matrices were separately tested against a predictor matrix reflecting a perfect sigmoidal model using GLM. The two resulting arrays of t-values were correlated across participants (right: example data for illustrative purpose). The same analysis was also performed using a perfect linear predictor matrix (see Figure 4).

Supplemental Experimental Procedures

Participants. 25 participants took part in this study of which 3 participants were excluded from further data analyses due to excessive head motion during scanning (high occurrence of >5mm movement and >3 degrees rotation spikes during a session). In addition, 2 participants were excluded due to scanner malfunctioning. Of the remaining 20 participants, 6 were male and 14 female, with an average age of 21.3 years (ranging between 19 and 28). Participants gave written consent and were paid for participating, as approved by the local Research Ethics Committee (CMO region Arnhem-Nijmegen, The Netherlands). All participants had normal or corrected-to-normal vision and reported to be in good health.

Virtual reality environments. We developed a virtual reality (VR) task using the UnrealEngine2 Runtime software (http:/udn.epicgames.com/Two/Webhome.html). Within this software we created a virtual arena that comprised a circular grassy plane surrounded by a wall with mountains in the background. By systematically changing the shape of the mountains we created six different environments (A through F). Since hippocampal fMRI responses are sensitive to changes in environmental geometry [S1-2], we shaped the mountains such that they linearly morphed from A through F (B = 20%A + 80%F; C = 40%A + 60%F; D = 60%A + 40%F; E = 80%A + 20%F.; see Figure 1B). The backgrounds of the two extreme environments (A and F) were chosen such that they are clearly discriminable but differ in as few features as possible to allow for smooth morphing between them. This resulted in both these backgrounds containing two prominent mountains of different heights and the same feature-restrained sky. Background images were created by generating height maps using Matlab 7.9 (http://www.mathworks.com) which were transformed into images of landscapes using the Terragen software package (http://www.planetside.co.uk). In the VR environment participants embodied an avatar with first person perspective and moved by using a button-box with their right hand enabling forward movement and left and right turns. The avatar's location and heading were recorded every 100 milliseconds.

Virtual reality task. Our task consisted of a two day learning phase (separated by 23 hours) immediately followed by a testing phase. In the learning phase participants learned the location of four objects in two separate environments, each object having a unique locations in each environment (Figure 1A). In the testing phase participants replaced the objects in both the learned environments (A and F) and the four novel morph environments (B, C, D and E).

The learning phase started with a short initial familiarization phase in which the four objects were shown once in environment A and once in F. In each subsequent trial an object was cued by appearing in the top part of the screen for 2 seconds after which participants navigated to the location they thought the object should be (time-out after 20 seconds, mean trial duration = 14.63 seconds). Feedback was provided by showing the object at its correct position after which participants had to collect the object (mean duration = 6.42 seconds; see Figure 1C). Every 9 minutes environments were switched by spawning a navigable bridge connecting the two environments that the participants had to cross. Participants visited each environment 3 times on each training day (first day: mean number of trials = 159.0, SEM = 6.8. second day: mean number of trials = 149.7, SEM = 8.6). In the testing phase participants performed between 1 and 4 trials before an ITI of 2-4 seconds appeared. Unknown to the participants, environments were changed during an ITI to one of the six environments in the morph sequence. Trial structure for the testing phase was unchanged other than that no feedback on the true object location was provided (mean number of trials = 235.5, SEM = 7.7, mean trial duration = 12.38 seconds). The order of environments and objects was counterbalanced so that no object would be cued twice in a row and the unique object-environment combinations are distributed uniformly over a session. In addition, for 15 participants we recorded a confidence rating (on a 5 point scale) on the precision of placement of the object after each trial.

Analysis of behavioral data. Spatial memory performance during all sessions was measured as the Euclidean distance between the response location and the correct object location in each trial. In order to test if participants' spatial memories were stable after the learning session, we fitted this 'drop error' over time of both training sessions to a quadratic polynomial model. The differential of the resulting function determined the slope at any given point on the learning curve. An average negative slope indicated learning of the location of the objects while the existence of a horizontal asymptote indicated maximum acquisition on a given day (Figure S1).

During the testing phase, our main behavioral measure was the relative difference between the object replacement location and the true object location in environments A (ΔA) and F (ΔF), calculated as $\Delta A/(\Delta A+\Delta F)$. This behavioral similarity measure scales linearly from 0 to 1 with increasing ΔA and decreasing ΔF (Figure 1D), reflecting the 'A-ness' and 'F-ness' of each memory response. This measurement was fitted to a sigmoid and a linear model using Maximum Likelihood

Estimation (MLE), for which the probability density functions were defined as six normal distributions with the mean of each distribution corresponding to values predicted by variable parameters for a linear or a sigmoid model. The standard deviation of all normal distributions was fixed to 0.25 so that at least 95% of the possible values would fall within the probability distribution. The resulting probability values were normalized so the area under the curve of the probability distributions summed to 1. Both the linear and the sigmoidal model have 2 free parameters, thus model complexity was held constant. The sigmoidal model was defined as

$$y_i = \frac{1+a}{1+e^{-x_i+b}+(1+a)/2}$$

where a is the amplitude and b the horizontal offset. The linear model was defined as

$$y_i = a * x_i + b$$

where a is the slope and b the vertical offset. Per participant, we fitted each model to the behavioral data and compared the resulting residual sum of square (RSS) between the models using a paired t-test.

Behavioral control experiment - Perception of background cues. We performed a separate behavioral experiment to test if there was non-linearity in the perception of the environments' background images (from A to F). 16 naive participants (12 females and 4 males, average age: 22.0 years; range: 16 to 29 years) were presented with 392 trials each consisting of image pairs representing the background of two environments. The images were displayed consecutively (stimulus duration lasted 2 seconds for each image) with a 1 second mask image in between (scrambled version of the first image of each pair). At the end of each trial, participants scored the perceived difference between the presented images on a 5-point similarity scale, followed by an ITI of 1 second, see Figure 2A. Trials were divided into 8 blocks separated with 20 seconds breaks. We applied to the similarity scores the same analysis logic as to the behavioral memory responses from the main experiment, i.e. the scored difference of each environment compared to either environment A (ΔA) or environment F (ΔF) represented as ΔA/(ΔA+ΔF). This measure was fitted to a sigmoid and linear model using MLE (see above). These models were then compared using a paired t-test on the resulting RSS.

MRI data acquisition. Imaging data was acquired on a Siemens 3T Trio scanner using a 32- channel head coil. The functional sequence used was a custom multi-echo 3D EPI sequence (TR = 1800 ms; TE = 25 ms; flip angle = 15°; 64 slices of matrix 112 × 112 with a 25% gap; voxel size = 2 × 2 × 2 mm). Scanning of functional images was subdivided into two blocks of 30 minutes each (1000 volumes each) with a break of approximately five minutes. In addition to the testing session, 11 out of the 20 participants were scanned during the first training session. For every scanning session, a field map using a gradient echo sequence was recorded for distortion correction of the acquired EPI images. The structural scan comprised a MPRAGE-grappa sequence (TR = 2300 ms; TE = 3.03 ms; flip angle = 8°; in-plane resolution = 256x256 mm; number of slices = 192; acceleration factor PE = 2; voxel resolution = 1 mm^3, duration = 321 seconds).

Image preprocessing. For the fMRI analysis, we used the Automatic Analysis framework [S3], which combines tools from SMP8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/), FreeSurfer (http://surfer.nmr.mgh.harvard.edu/) and the FMRIB Software Library v5.0 (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/), complemented by custom scripts executed in Matlab 7.9. Preprocessing consisted of de-noising of the structural image using a non-local means filter [S4], motion correction, co-registration of functional images to the structural scan and brain-extraction. White matter was masked out using FreeSurfer. Finally, a study specific structural template was build and non-linear normalization of participant-specific contrasts to this template was performed using the Advanced Normalization Tools version 1.9 for Unix (http://www.picsl.upenn.edu/ANTS).

Representational similarity analysis. fMRI data were modeled with square wave regressors, one regressor for each trial. The start of each trial was defined by the onset of a cue indicating which object should be relocated by the participant and the response by the participant defined the end of each trial. The average trial duration was 14.63 seconds (i.e. 8.1 TRs). The regressors were convolved with a hemodynamic response function (HRF) and voxel-wise unsmoothed beta values were extracted per regressor of interest. Additional nuisance regressors consisted of three translations and three rotations derived

from motion correction, as well as single spike regressors reflecting large, sudden, head movements. The voxel-wise betas were used in subsequent searchlight analyses using representational similarity analysis (RSA) [S5]. RSA comprised the voxel-wise extraction of beta values from each trial, which were correlated across trials resulting in a trial by trial correlation matrix. For every combination of trials a prediction on their similarity was made based on the environments navigated in those trials (Figure 3A). How well the similarity matrices fit the prediction models was assessed using general linear modeling (GLM). The resulting searchlight voxel-wise beta brain maps were normalized, smoothed and subjected to second-level random effects analysis.

Group-level fMRI analysis. In a first analysis on fMRI data from the testing phase (see Figure 3), the sigmoid fitted to each participant's behavioral similarity measures $\Delta A/(\Delta A+\Delta F)$ was used to predict the similarity in the multi-voxel fMRI pattern on a trial-by-trial basis. The resulting predictor matrix (see Figure S4B) was subsequently tested against the actual similarity between different trials from the fMRI data. A linear model was used as control since this adhered to the visual changes in the environment morph sequence (Figure 2C; see Behavioral control experiment above).

Given our strong a prior hypothesis, analyses were initially restricted to the hippocampus, details on the region-of-interest are outlined below. Using a bootstrap method we applied a corrected statistical threshold of $P<0.05$ to all GLM results. A bootstrap method was also used to directly compare the fit between these models while keeping multiple comparison correction (see Bootstrap analysis below). To assess the relationship between putative attractor dynamics in both behavior and neural pattern (see Figure 4), we compared how well a perfect sigmoidal model (step-function) fit the behavioral and fMRI data, using again a linear model as control (Figure S4A). Fitting of models to both the behavioral data ($\Delta A/(\Delta A+\Delta F)$ and fMRI data (multi-voxel trial- by-trial fMRI pattern) was performed using GLM. Subsequently, resulting t-maps were correlated over participants, separately for sigmoidal and linear models (Figure 4B). Again, a bootstrap method was used to assess the significance of the correlations while keeping multiple comparison correction (see Bootstrap analysis below).

Bootstrap analysis. To assess whether fMRI data in any voxel significantly fit a sigmoidal or linear model we shuffled all the first level contrast images and applied a 5 mm half width half maximum variance smoothing to the contrasts. We did this 5000 times and for each permutation the resulting data was tested against the sigmoidal and linear prediction model generated as if the data was non-shuffled. The resulting distribution of t-values were used for testing the significance of fit of either model to non-shuffled data. A similar bootstrap method was used to directly compare the fit between these models while keeping multiple comparison correction. Per participant we subtracted the GLM t-value of the sigmoidal model from the t-values of the linear model and using a one-sample t-test on this difference a p-value was calculated. This step was repeated for 5000 permutations, each time shuffling the voxels, and the resulting distribution of p-values provided the critical value to test the p-value of a one-sample t-test on non-shuffled data. A bootstrap method was also used to assess where in the hippocampus significant correlation between fMRI data and behavioral data existed. This method comprised shuffling the t-values obtained by applying a GLM on fMRI data using either a perfect sigmoidal or linear model over participants for every voxel. The resulting t-values were correlated with the t-values resulting from fitting the same models to non-shuffled behavioral data. This process was repeated 5000 times and rendered a distribution of correlation coefficients per voxel used to test the significance of the correlation coefficient between the non-shuffled fMRI and behavioral data.

Regions Of Interest. A hippocampal ROI was manually segmented using MeVisLab software (MeVis Medical Solutions AG, Bremen, Germany) on a study specific structural template generated by the ANTS software (see above). The segmentation was based on a protocol described by [S6] and [S7], and guided by an anatomical atlas of the human hippocampus [S8]. Post-hoc ROI analyses using the perfect sigmoidal and linear model were performed on the perirhinal, entorhinal and parahippocampal cortex. For the perirhinal cortex ROI, a probabilistic map was used [S9]. The entorhinal cortex ROI [S10] was obtained from the SPM anatomy toolbox version 2.2 and the parahippocampal cortex ROI was obtained from the Talairach label database [S11-12] and transformed into MNI space using BioImage Suite version 3.01 (http://bioimagesuite.yale.edu). All probabilistic maps were visually inspected to ensure that no hippocampal voxels were included and a relative conservative threshold was applied (voxels included were in >90% of brains used to make the probabilistic maps labeled as the ROI).

Searchlight analysis. Searchlight mapping was performed on the native space images of each participant by moving a spherical searchlight (4 voxel radius) through the gray-matter masked volume or ROI one voxel at a time. Statistics were mapped back to the central voxel of each spherical searchlight thus yielding a single-participant information map. Analysis

was restricted to searchlights that contained at least 30 voxels, thereby eliminating searchlights that were close to the edge of gray matter. The first-level results were normalized and smoothed with a 4mm FWHM kernel, and a second-level model on the resulting data was carried out to examine information at the group level. All results are reported in MNI coordinates.

Divergence of the neural pattern between environments during training. In a post-hoc analysis on the training phase data, we tested if the neural pattern in the hippocampus changed as a function of learning object positions while navigating environments A and F. We split the data into two halves (first and last 30 minutes of the session) and calculated per trial the voxel-wise correlations with every other trial within each half. The resulting correlations were related to the appropriate cell of an environment-by-environment correlation matrix and this matrix was tested against a matrix predicting a larger neural similarity for activation patterns within an environment than between environments (Figure S1). We predicted that the model would fit the second half of the learning phase better than the first half as the representation of the environments was thought to diverge over time, consistent with attractor dynamics.

Monte-Carlo simulation. In order to test if a sigmoidal model fits the fMRI data better regardless of the specific environment navigated or the trial number, we performed a Monte-Carlo simulation. From the whole data set each voxel from each EPI image was assigned a unique number. From this pool 105750 voxels were randomly picked with replacement and assigned a trial number between 1 and 235 and an environment number between 1 and 6 such that each combination of trial and environment had the same amount of voxels. Using this data, we fitted both a sigmoidal and a linear model as described above and saved the resulting t-values. The process of picking voxels, assigning trials and scenes and fitting both models was repeated 10,000 times resulting in a distribution of t-values for each model, which were compared using the non-parametric Kolmogorov-Smirnov test.

fMRI and behavioral data similarity. To examine the relation between behavioral and fMRI data we calculated a similarity measure for the fMRI data akin to the similarity measure used to analyze the behavioral data. Using GLM, for each environment we modeled both similarity to environment A (ΔA) and similarity to environment F (ΔF). Using the peak voxel from our correlation analysis (see Figure 4A) as our coordinate, we applied the formula $\Delta A/(\Delta A+\Delta F)$ which resulted in a similarity measure that scales linearly from 0 to 1 with increasing ΔA and decreasing ΔF. Having both a similarity measure from behavioral and fMRI data that can be readily compared, we plotted these measures against each other and applied K-means clustering on the resulting 2D space, assessing the optimal number of clusters using the Calinski-Harabasz index value [S13], and examined the correlation between the measures using Pearson's correlation coefficient. K-means clustering was performed 1000 times, each time with new random starting centroids to avoid convergence to a local minima and the result with the lowest total within-cluster point-to-centroid distances is shown in Figure 4C.

Within-trial dynamics of sigmoidacity versus linearity. In a post-hoc analysis we looked at within-trial dynamical changes of the explanatory power of both a perfect sigmoidal and linear model over time, again around the peak voxel obtained from our correlation analysis (see Figure 4A). To this end, we split each trial in overlapping 3.6 second segments with the center of the segments 1.8 seconds apart from neighboring segments. Using GLM, all segments that were in the same time-window within all trials were fitted to a perfect sigmoidal model and the resulting betas were divided by betas obtained from fitting the same data to a linear model. This resulted in a measure directly comparing the contribution of the sigmoidal model to the linear model as trials progressed. This was done for each participant, resulting in no more than 20 data-points per time-window. Using a linear regression we assessed whether this measure changed structurally within trials as they progressed.

Supplemental References

S1. Bird, C. M., Capponi, C., King, J. A., Doeller, C. F. and Burgess, N. (2010). Establishing the boundaries: the hippocampal contribution to imagining scenes. J. Neurosci. *30*, 11688-11695.

S2. Stokes, J., Kyle, C. and Ekstrom, A. D. (2015). Complementary roles of human hippocampal subfields in differentiation and integration of spatial context. J. Cogn. Neurosci. *27*, 546-559.

S3. Cusack, R., Vicente-Grabovetsky, A., Mitchell, D. J., Wild, C. J., Auer, T., Linke, A. C. and Peelle, J. E. (2014). Automatic analysis (aa): efficient neuroimaging workflows and parallel processing using Matlab and XML. Front.

Neuroinform. *8*, 90.

S4. Coupe, P., Yger, P., Prima, S., Hellier, P., Kervrann, C. and Barillot, C. (2008). An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. IEEE. Trans. Med. Imaging. *27*, 425-441.

S5. Kriegeskorte, N., Mur, M. and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4.

S6. Malykhin, N. V., Lebel, R. M., Coupland, N. J., Wilman, A. H. and Carter, R. (2010). In vivo quantification of hippocampal subfields using 4.7 T fast spin echo imaging. Neuroimage *49*, 1224-1230.

S7. Wisse, L. E. M., Gerritsen, L., Zwanenburg, J. J. M., Kuijf, H. J., Luijten, P. R., Biessels, G. J. and Geerlings, M. I. (2012). Subfields of the hippocampal formation at 7 T MRI: in vivo volumetric assessment. Neuroimage *61*, 1043-1049.

S8. Duvernoy H. M., Cattin F. and Risold P. Y. (2013). The Human Hippocampus: Functional Anatomy, Vascularization and Serial Sections with MRI (Berlin: Springer)

S9. Holdstock, J. S., Hocking, J., Notley, P., Devlin, J. T. and Price, C. J. (2009). Integrating visual and tactile information in the perirhinal cortex. Cereb Cortex *19*, 2993-3000.

S10. Amunts, K., Kedo, O., Kindler, M., Pieperhoff, P., Mohlberg, H., Shah, N. J., Habel, U., Schneider, F. and Zilles, K. (2005). Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: intersubject variability and probability maps. Anat Embryol (Berl) *210*, 343-352.

S11. Lancaster, J. L., Rainey, L. H., Summerlin, J. L., Freitas, C. S., Fox, P. T., Evans, A. C., Toga, A. W. and Mazziotta, J. C. (1997). Automated labeling of the human brain: a preliminary report on the development and evaluation of a forward-transform method. Hum Brain Mapp *5*, 238-242.

S12. Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., Kochunov, P. V., Nickerson, D., Mikiten, S. A. and Fox, P. T. (2000). Automated Talairach atlas labels for functional brain mapping. Hum Brain Mapp *10*, 120-131.

S13. Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis Communications in Statistics-theory and Methods *3*, 1-27.