Medial prefrontal cortex compresses concept representations through learning

Michael L. Mack
Department of Psychology
University of Toronto
Toronto, ON, Canada
mack@psych.utoronto.ca

Alison R. Preston
Center for Learning and Memory
The University of Texas at Austin
Austin, TX, USA
apreston@utexas.edu

Bradley C. Love
Experimental Psychology
University College London
London, UK
b.love@ucl.ac.uk

Abstract— **Prefrontal cortex (PFC) is thought to support the ability to focus on goal-relevant information by filtering out irrelevant information, a process akin to dimensionality reduction. Here, we find direct evidence of goal-directed data compression within medial PFC during learning, such that the degree of neural compression predicts an individual's ability to selectively attend to concept-specific information. These findings suggest a domain-general mechanism of learning through compression in mPFC.**

***Keywords*— prefrontal cortex; fMRI; attention; category learning; computational modeling**

I. Introduction

Prefrontal cortex (PFC) is sensitive to the complexity of incoming information [1] and theoretical perspectives suggest that a core function of PFC is to focus representation on goal-relevant features by filtering out irrelevant content [2], [3]. In particular, medial PFC (mPFC) is thought to represent the latent structures of experience [4], [5], coding for causal links [6] and task-related cognitive maps [7]. At the heart of these accounts is the hypothesis that during learning, mPFC performs data reduction on incoming information, compressing features that do not matter to emphasize encoding of goal-relevant information structures. Although emerging evidence suggests structured representations occur in the rodent homologue of mPFC [8], such coding in human PFC remains poorly understood. Here, we directly assess the data reduction hypothesis by leveraging an information-theoretic approach in human neuroimaging to measure how learning is supported by mPFC compression processes.

II. Methods

A. Participants

Twenty-three volunteers (11 females, mean age 22.3 years old, ranging from 18 to 31 years) participated in the experiment. All subjects were right handed, had normal or corrected-to-normal vision, and were compensated $75 for participating. One participant did not perform above chance in one of the learning problems, thus was excluded from analysis.

B. Stimuli

Eight color images of insects were used in the experiment (Fig. 1A). The insect images consisted of one body with different combinations of three features: legs, mouth, and antennae. There were two versions of each feature (thick or thin antennae, thick or thin legs, and shovel or pincer mandible). The eight insect images included all possible combinations of the three features.

C. Procedures for learning problems

We focused on concept learning, given the recent findings that mPFC represents conceptual information in an organized fashion [9]. Participants learned to classify the same insect images across three different problems [10] (Fig. 1A). These learning problems were defined by rules for which a different number of features had to be consider to successfully classify (see Table 1): the low category complexity problem was unidimensional (e.g., insects living in warm climates have thick legs, cold climate insects have thin legs), the medium category complexity problem depended on two features (e.g., insects from rural environments have thick antennae and shovel mandible or thin antennae and pincer mandible, urban insects have thick antennae and pincer mandible or thin antennae and shovel mandible), and the high category complexity problem required all three features (i.e., each insect's class was uniquely defined by a combination of features).

This design allowed us to manipulate the complexity of the conceptual space needed represent each problem (see Fig 1A). Complexity and compression have an inverse relationship; the lower the complexity of a conceptual space, the higher the degree of compression. For instance, in learning the unidimensional problem, variance along the two irrelevant

978-1-5386-3159-1/17/$31.00 ©2017 IEEE

feature dimensions can be compressed resulting in a lower complexity conceptual space. In contrast, learning the high complexity problem requires less compression because all three feature dimensions must be represented, resulting in a relatively more complex conceptual space.

TABLE I. STIMULUS FEATURES AND CLASS ASSOCIATIONS

	feature attribute			**problem complexity**		
stimulus	1	2	3	low	medium	high
1	0	0	0	A	A	B
2	0	0	1	A	B	A
3	0	1	0	A	B	A
4	0	1	1	A	A	B
5	1	0	0	B	A	A
6	1	0	1	B	B	B
7	1	1	0	B	B	B
8	1	1	1	B	A	A

Differences in complexity across the three learning problems thus provide a means for testing how learning shapes the dimensionality of neural concept representations. Namely, brain regions involved in data compression should *learn* to represent less complex problems with fewer dimensions. To test this prediction, we recorded functional magnetic resonance imaging (fMRI) data while participants learned the three problems and measured the degree that multivoxel activation patterns were compressed through learning using principal component analysis (PCA; Fig 1B), a method for low-rank approximation of multidimensional data [11].

III. Results

A. *Neural compression*

We assessed the representational complexity of the neural measures of stimulus representation during learning with a searchlight method [13]. Using a searchlight sphere with a radius of 4 voxels (voxels per sphere: 242 mean, 257 mode, 76 minimum, 257 maximum), we extracted a vector of activation values across all voxels within a searchlight sphere for all 32 trials within a problem run. These activation vectors were then submitted to PCA to assess the degree of correlation in voxel activation across the different trials. PCA was performed using the singular value decomposition method as implemented in the *decomposition.PCA* function of the scikit-learn (version 0.17.1) Python library. To characterize the amount of dimensional reduction possible in the neural representation, we calculated the number of principal components that were necessary to explain 90% of the variance (k) in the activation vectors. We scaled this number into a compression score, *1-k/n*, where n is equal to 32, the total number of activation patterns submitted to PCA. By definition, 32 PCs will account for 100% of the variance, but no compression. With this definition of neural compression, larger compression scores indicated fewer principal components were needed to explain the variance across trials in the neural data (i.e., neural representations with lower dimensional complexity). In contrast, smaller compression scores indicated more principal components were required to explain the variance (i.e., neural representations with higher dimensional complexity). This neural compression searchlight was performed across the whole brain separately for each participant and each run of the three learning problems in native space.

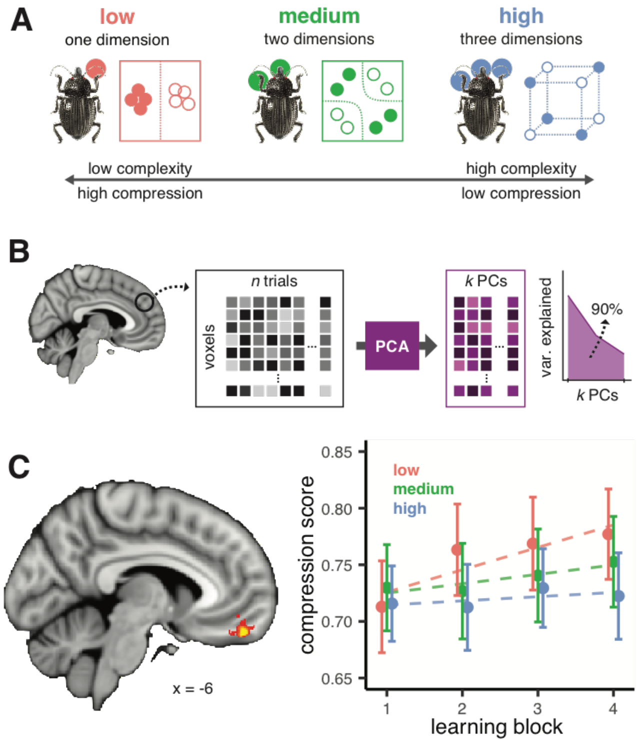

Fig. 1. Experimental schematic and neural compression analysis. A) The learning problems differed in rule complexity (see Table 1). Low complexity was unidimensional (e.g., antennae size), medium complexity required a conjunction of two features (e.g., leg size and mandible shape), and high complexity required all three features. B) Principal component analysis (PCA) was performed on neural patterns evoked for each of n trial within a learning block. The number of principal components (PC) required to explain 90% of the variance (k) was used to calculate a neural compression score (1-k/n). We quantified neural compression as a function of learning problem and repetition; this interaction reflects changes in the complexity of neural representations that emerge with learning. C) A wholebrain searchlight revealed an mPFC region that showed an interaction between learning block and problem complexity (i.e., compression for low > medium > high). Error bars represent 95% confidence intervals. (N=22).

Throughout the entire brain, only mPFC showed the predicted relationship between compression and conceptual complexity (peak F=14.4; peak coordinates -5, 51, -20; 1257 voxels; Fig. 1C). Importantly, mPFC compression emerged over learning blocks ($F_{6,126}$=3.27, MSE=0.002, p=0.005, η_p^2=0.135). Because the stimuli were identical across the three problems, this finding demonstrates that learning-related compression is goal-specific, with mPFC requiring fewer dimensions for less complex goals.

B. *Relating neural compression to selective attention*

To assess whether mPFC compression tracked changes in attentional allocation, we characterized the participant-specific attentional weights given to each stimulus feature across the

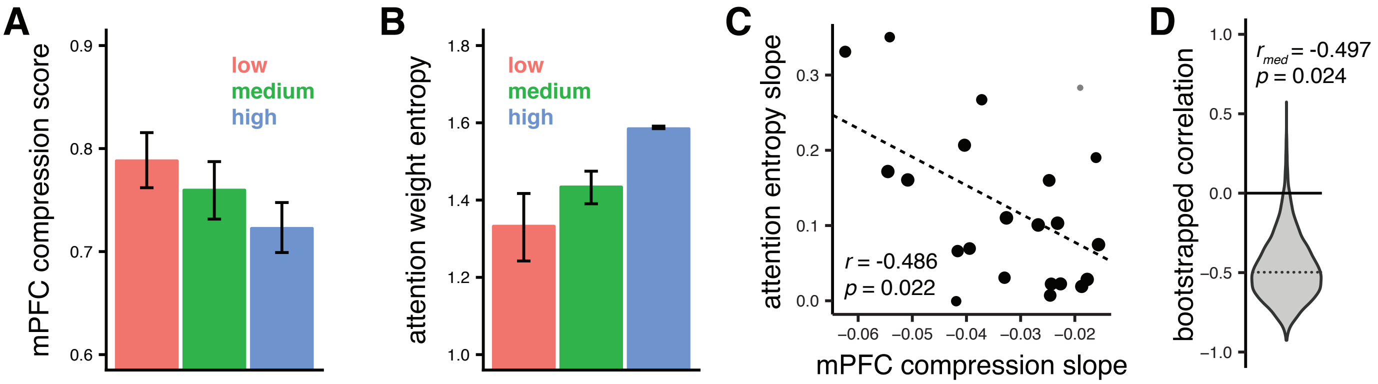

Fig. 2. Relationship between mPFC compression and model-based attention weighting (*N*=22). **A)** mPFC neural compression decreased across problems, consistent with the complexity demands. **B)** Attention weight entropy (i.e., dispersion in attention weights) mirrored neural compression, showing attentional strategies consistent with the feature relevancy across the problems. Error bars represent 95% CI. **C)** Changes in mPFC compression (indexed as the slope of compression across problems) predicted the degree of problem-specific attention weighting (indexed as the attention entropy slope) across participants. The size of the scatterplot points depicts the weighting from a robust regression analysis; the dashed line depicts the best-fitting regression line. **D)** A bootstrapping procedure confirmed the relationship between neural compression and attention entropy.

three problems using a computational learning model, SUSTAIN [14]. SUSTAIN is a network-based learning model that classifies incoming stimuli by comparing them to memory-based knowledge representations of previously experienced stimuli. Sensory stimuli are encoded by SUSTAIN into perceptual representations based on the value of the stimulus features. The values of these features are biased according to attention weights operationalized as receptive fields on each feature attribute. During the course of learning, these attention weight receptive fields are tuned to give more weight to diagnostic features. SUSTAIN represents knowledge as clusters of stimulus features and class associations that are built and tuned over the course of learning. New clusters are recruited and existing clusters updated according to the current learning goals. A full mathematical formulization of SUSTAIN is provided in its introductory publication [14].

To assess attentional strategy during the learning problems, we fit each participant's learning performance with SUSTAIN and extracted the attention weights at the end of learning in each problem. We then calculated the entropy across the attention weights for each problem. Attention weight entropy indexed changes in attentional allocation; high entropy indicates equivalent weighting to all three features, whereas low entropy indicates attention directed to only one feature. We found that across the learning problems, attention weight entropy increased with conceptual complexity (χ^2_2=33.17, p=6.26×10^{-8}; Fig. 2B). Importantly, the increase in attention weight entropy mirrored the decrease in mPFC neural compression (χ^2_2=24.82, p=4.08×10^{-6}; Fig. 2A), suggesting a link between the behavioral and neural signatures of dimensionality reduction.

To assess this relationship, we evaluated whether participants' attention weights were predicted by mPFC neural compression at the individual participant level. We first calculated the degree of change in neural compression and attention weight entropy across the three problems. We did this by fitting separate linear regression models to each participant's neural compression scores (average compression score within the participant-specific mPFC cluster mask as described above) and attention weight entropy values. This resulted in neural compression and attention entropy slopes for each participant (plotted in Fig. 2C). Specifically, for neural compression, a more negative slope reflects decreasing compression (and increasing representational complexity) across the low, medium, and high complexity problems. A flat slope would suggest no change in representational complexity across the problems. For attention weight entropy, a more positive slope reflects increasing entropy in attention weights across low, medium, and high problems. Such a slope would be found when attention is optimally deployed across the three problems: one feature in low, two features in medium, and all three features in high. A flat slope would suggest attention was equivalently deployed across the problem types.

If the ability to compress neural representations in a problem-appropriate fashion is related to participants' ability to attend to problem-relevant features, the prediction follows that participants with changes in neural compression across problem (i.e., more negative neural compression slopes) will also show the greatest change in selective attention (i.e., more positive attention entropy slopes). Our analysis confirmed this hypothesis (r_{20}=-0.486, p=0.022; see Fig. 2C).

To assess the reliability of this finding and evaluate the influence of potential outliers, we performed three additional analyses. First, we analyzed the relationship between neural compression and attention entropy with robust regression using a logistic weighting function. Robust regression accounts for potential outlier observations by down weighting observations that individually influence the estimation of a linear regression model between two variables. Consistent with the correlation results, the robust regression results showed evidence of a linear relationship between neural compression and attention weight entropy (β = -4.019, SE = 1.657, t = -2.427, p = 0.025). The weighting of each observation estimated in the robust regression analysis is depicted in Figure 2C as the relative size of the data points. Second, we identified and removed potential outliers by evaluating the standardized difference in fit statistic (DFFITS) for each observation. Using the standard DFFITS threshold [15], one observation was identified as an outlier (noted as a grey data point in Fig. 2C). This observation was excluded from a linear regression analysis between neural compression and attention weight entropy. The results of this analysis showed that even with this potential outlier observation removed a strong relationship remained (β = -4.568, *SE* = 1.379, *t* = -3.312, *p* = 0.004, R^2 = 0.366). Third, we performed a nonparametric bootstrap analysis to assess the robustness of the correlation between neural compression and attention entropy. We randomly sampled with replacement from the slope pair observations 5000 times, calculating and storing the Pearson correlation coefficient on each iteration. The resulting distribution of correlation coefficients revealed a significant relationship (see Fig. 2D; median *r* = -0.497, *p = 0.024, 95% CI* [-0.788, -0.002]).

In a final analysis, we assessed the direct relationship between neural compression and attention weight entropy. To do this, we performed a mixed effects linear regression using the *lme4* package (version 1.1-12) in *R* (version 3.3.2). The mixed effects model was defined such that the attention weight entropy values were directly predicted by mPFC neural compression with participants as a random factor. Consistent with the slope-based analyses, the direct model revealed a significant relationship between neural compression and attention entropy (β=-0.776, *SE*=0.329, *t*=-2.358, *p*=0.024). We confirmed this effect with a bootstrap procedure to generate a distribution of regression coefficients, β_{boot}. This consisted of 5000 iterations in which participants were randomly sampled with replacement and the regression model was re-estimated. The results of this analysis demonstrated a robust relationship between mPFC compression and attention entropy (*median* β_{boot}=-0.770, *p*=0.011, *95% CI* [-1.431, -0.117]). Collectively, these findings suggest that the degree of problem-specific neural compression in mPFC predicted participants' attentional strategies.

IV. Discussion

By focusing on a mechanism by which mPFC forms and represents concepts through goal-sensitive dimensionality reduction, we show that activation patterns within mPFC code for the complexity of concepts. Critically, by evaluating behavior through the lens of a computational model, we also demonstrate that concept-specific mPFC coding is related to learning. These findings support the view that mPFC builds cognitive maps [7], [9], [16], structuring representations to highlight goal-specific features and compress irrelevant information. Such a mechanism could be critical for many processes associated with mPFC including schema representation [17], latent casual models [7], grid-like conceptual maps [9], and value coding [18], [19].

Acknowledgments

Thanks to Christiane Ahlheim for manuscript comments. M.L.M. was supported by NIMH grant F32-MH100904; A.R.P. by NIMH grant R01-MH100121, and NSF CAREER Award 1056019; and B.C.L by Leverhulme Trust grant RPG-2014-075, Wellcome Trust Senior Investigator Award WT106931MA, and NICHD grant 1P01HD080679.

References

[1] D. Badre, A. S. Kayser, and M. D'Esposito, "Frontal cortex and the discovery of abstract action rules," *Neuron*, vol. 66, no. 2, pp. 315–326, 2010.

[2] R. C. Wilson, Y. K. Takahashi, G. Schoenbaum, and Y. Niv, "Orbitofrontal cortex as a cognitive map of task space," *Neuron*, vol. 81, no. 2, pp. 267–278, 2014.

[3] V. Mante, D. Sussillo, K. V Shenoy, and W. T. Newsome, "Context-dependent computation by recurrent dynamics in prefrontal cortex.," *Nature*, vol. 503, no. 7474, pp. 78–84, 2013.

[4] D. Zeithamova, A. L. Dominick, and A. R. Preston, "Hippocampal and Ventral Medial Prefrontal Activation during Retrieval-Mediated Learning Supports Novel Inference," *Neuron*, vol. 75, pp. 168–179, 2012.

[5] M. L. Schlichting, J. A. Mumford, and A. R. Preston, "Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex," *Nat. Commun.*, vol. 6, p. 8151, Aug. 2015.

[6] S. C. Y. Chan, Y. Niv, and K. A. Norman, "A Probability Distribution over Latent Causes, in the Orbitofrontal Cortex.," *J. Neurosci.*, vol. 36, no. 30, pp. 7817–28, 2016.

[7] N. W. Schuck, M. B. Cai, R. C. Wilson, Y. Niv, N. W. Schuck, M. B. Cai, R. C. Wilson, and Y. Niv, "Human Orbitofrontal Cortex Represents a Cognitive Map of State Space Article Human Orbitofrontal Cortex Represents a Cognitive Map of State Space," *Neuron*, vol. 91, no. 6, pp. 1402–1412, 2016.

[8] A. Farovik, R. J. Place, S. McKenzie, B. Porter, C. E. Munro, and H. Eichenbaum, "Orbitofrontal Cortex Encodes Memories within Value-Based Schemas and Represents Contexts That Guide Memory Retrieval," *J. Neurosci.*, vol. 35, no. 21, pp. 8333–8344, 2015.

[9] A. O. Constantinescu, J. X. O'Reilly, and T. E. J. Behrens, "Organizing conceptual knowledge in humans with a gridlike code," *Science (80-.).*, vol. 352, no. 6292, pp. 1464–1468, 2016.

[10] R. N. Shepard, C. I. Hovland, and H. M. Jenkins, "Learning and memorization of classification," *Psychol. Monogr.*, vol. 75, no. 13, p. 517, 1961.

[11] C. Eckart and G. Young, "The Approximation of One Matrix by Another Low Rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[12] J. A. Mumford, B. O. Turner, F. G. Ashby, and R. A. Poldrack, "Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses," *Neuroimage*, vol. 59, no. 3, pp. 2636–2643, 2012.

[13] N. Kriegeskorte, R. Goebel, and P. Bandettini, "Information-based functional brain mapping," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 10, pp. 3863–3868, 2006.

[14] B. C. Love, D. Medin, and T. M. Gureckis, "SUSTAIN: A network model of category learning," *Psychol. Rev.*, vol. 111, no. 2, pp. 309–332, 2004.

[15] H. Aguinis, R. K. Gottfredson, and H. Joo, "Best-Practice Recommendations for Defining, Identifying, and Handling Outliers," *Organ. Res. Methods*, vol. 16, no. 2, pp. 270–301, 2013.

[16] A. M. Wikenheiser and G. Schoenbaum, "Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex," *Nat Rev Neurosci*, vol. 17, no. 8, pp. 513–523, 2016.

[17] M. T. R. Van Kesteren, D. J. Ruiter, G. Fernández, and R. N. Henson, "How schema and novelty augment memory formation," *Trends Neurosci.*, vol. 35, no. 4, pp. 211–219, 2012.

[18] J. A. Clithero and A. Rangel, "Informatic parcellation of the network involved in the computation of subjective value," *Soc. Cogn. Affect. Neurosci.*, vol. 9, no. 9, pp. 1289–1302, 2013.

[19] M. Grueschow, R. Polania, T. A. Hare, and C. C. Ruff, "Automatic versus Choice-Dependent Value Representations in the Human Brain," *Neuron*, vol. 85, no. 4, pp. 874–885, 2015.