Core episodic encoding and retrieval processes revealed by
dynamics of neural activity

James E. Kragel,1 Youssef Ezzyat,1 Michael R. Sperling,2 Richard Gorniak,3 Gregory A. Worrell,4 Brent M. Berry,4 Cory Inman,5 Jui-Jui Lin,6 Kathryn A. Davis,7 Sandhitsu R. Das,7 Joel M. Stein,8 Barbara C. Jobst,9 Kareem A. Zaghloul,10 Sameer A. Sheth,11 Daniel S. Rizzuto,1∗ Michael J. Kahana1∗

1 Department of Psychology, University of Pennsylvania, Philadelphia PA 19104, USA
2 Department of Neurology, 3 Department of Radiology, Thomas Jefferson University Hospital, Philadelphia PA 19107, USA
4 Department of Neurology, Mayo Clinic, Rochester MN 55905, USA
5 Department of Neurosurgery, Emory School of Medicine, Atlanta GA 30322, USA
6 Department of Neurosurgery, University of Texas Southwestern, Dallas TX 75390, USA
7 Department of Neurology, 8 Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia PA 19104, USA
9 Department of Neurology, Dartmouth Medical Center, Lebanon NH 03756, USA
10 Surgical Neurology Branch, National Institutes of Health, Bethesda MD 20814, USA
11 Department of Neurosurgery, Columbia University Medical Center, New York NY 10032, USA
∗These authors contributed equally to this work.

Running title: Spectral biomarkers of memory

Corresponding Author:
Michael J. Kahana
University of Pennsylvania, Department of Psychology
Suite 263
Stephen A. Levin Building
Philadelphia, PA 19104
e-mail: kahana@psych.upenn.edu
phone: (215) 746–3501
fax: (215) 746–6848

Conflict of Interest: The authors declare no competing financial interests.

Acknowledgments: We thank Blackrock Microsystems for providing neural recording equipment. This work was supported by the DARPA Restoring Active Memory (RAM) program (Cooperative Agreement N66001-14-2-4032), as well as National Institutes of Health grant MH55687. We are indebted to all patients who have selflessly volunteered their time to participate in our study. The views, opinions, and/or findings contained in this material are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.
Abstract

The capacity to remember the past is dependent on neural processing that occurs not only when an event is experienced but also when it becomes reinstated into awareness at a later point in time. Neural networks that span the medial temporal lobes (MTL), prefrontal cortex, and posterior cortical regions support the formation and retrieval of memories. However, the specificity of processing that occurs during learning and retrieval of encoded information is not well understood. To elucidate whether common or distinct processing during encoding and retrieval best predict memory performance, we administered a delayed free recall task to 187 neurosurgical patients implanted with subdural or intraparenchymal depth electrodes. We identified biomarkers of memory success that were jointly predictive of both successful memory formation and recall, in the form of increased high frequency activity in prefrontal, MTL, and inferior parietal cortices, accompanied by widespread decreases in low frequency power. These findings demonstrate that in the presence of distinct neural systems that support episodic encoding and retrieval of memories, variability in neural activity within a putative core memory network shapes the ability to remember the past.

Keywords: iEEG, MVPA, free recall, episodic memory
Introduction

Episodic memory is dependent upon learning operations that allow experiences to be encoded in distributed patterns of cortical activity across the brain, and retrieval operations that restate the contents of a memory when prompted by a retrieval cue (Polyn and Kahana, 2008). Theoretical models of episodic memory have been proposed wherein encoding and retrieval are supported by distinct neural substrates (Nyberg et al., 1996; Lepage et al., 1998; Kim, 2015). Neuroimaging studies provide evidence for these theoretical models, revealing dissociations between the neural correlates of episodic encoding and retrieval within the hippocampus (Zeineh et al., 2003; Eldridge et al., 2005), as well as functional networks spanning prefrontal, medial temporal, lateral temporal, and parietal cortical regions (Daselaar et al., 2009; Kim et al., 2010). Memory performance is also supported by neural processes that are general to both encoding and retrieval, as evidenced by studies demonstrating common patterns of activation within the hippocampus (Small et al., 2001; Stark and Okado, 2003; Prince et al., 2005) and across widespread cortical regions (Sederberg et al., 2007) during successful encoding and retrieval. Understanding the relative contributions of neural processes engaged during learning and retrieval is critical to understanding the mechanisms that ultimately determine whether information can be successfully retrieved at a later point in time. The goal of the present work is to determine whether memory performance is best predicted by distinct neural processes that occur during the encoding and retrieval of episodic information, or common electrophysiological states that may reflect the engagement of core mnemonic processes.

Theoretical models of episodic memory propose that successful memory retrieval involves the reinstatement of representations that were present during the initial experience of an event (McClelland et al., 1995; O’Reilly and Rudy, 2001). Using multivariate pattern analysis, it has been repeatedly demonstrated that patterns of cortical activation that are present during the encoding of an event are recapitulated during retrieval (Staresina et al., 2012; Ritchey et al., 2013; Wing et al., 2014). While considerable progress has been made linking the reinstatement of neural activity to the contents of retrieval (for a review, see Danker and Anderson 2010), the neural states that facilitate the encoding and reactivation of this content may differ. Multivariate classification techniques allow for the estimation of latent cognitive states, such as attentive states that improve task performance (Rosenberg et al., 2015), and incorporate information from neural populations across the brain. By training a multivariate classifier to identify neural signals that predict successful memory function, we can determine the relationship between the processes that mediate episodic encoding
and retrieval.

If the success of episodic memory is determined by the recapitulation of the same processes that were engaged during encoding (Kolers, 1973; Tulving and Thompson, 1973; Kolers and Roediger, 1984), one would predict that engagement of neural systems that are generally predictive of memory success (e.g., the hippocampus; Prince et al. 2005) should be able to identify when either successful encoding or retrieval operations occur. In contrast, if the function of neural systems that are specific to encoding or retrieval determine the fate of studied information, then classifiers trained to predict the success of memory formation should fail to identify when successful retrieval occurs. Recent electrophysiological investigations of human memory have demonstrated that shifts in the power spectrum, specifically increased high frequency activity (HFA) with concomitant decreases in low frequency activity (LFA) provide informative features for predicting successful memory formation (Burke et al., 2014a; Long et al., 2014; Greenberg et al., 2015) and retrieval (Burke et al., 2014b). These observations are consistent with intracranial studies of cortical function in humans that describe this phenomena as a marker of local cortical engagement (Manning et al., 2009; Miller et al., 2014; Burke et al., 2015b). Incorporating these measures of neural activity into multivariate classifiers enables the estimation of biomarkers of successful memory function, based on the patterns of activity of across distributed neuronal populations.

To determine whether the same neural mechanisms support the encoding and retrieval of episodic memories and to determine the degree to which these measures predict memory performance, we examined intracranial electroencephalographic (iEEG) data recorded from 187 patients undergoing clinical monitoring for seizures while they performed a free-recall task. We developed multivariate classifiers by applying L2-penalized logistic regression to patterns of spectral power associated with either successful or unsuccessful memory performance during the encoding and retrieval of verbal information. Using a cross-classification approach, we tested whether the same biomarkers predict successful memory formation and retrieval. Finally, we estimated general biomarkers of episodic memory function by constructing a joint classifier that incorporated information from both encoding and retrieval phases of the task. These analyses aimed to uncover the degree to which latent mnemonic states can be measured from brain activity, and the degree to which variation in patterns of spectral power common to both encoding and retrieval can predict memory function.
Materials and Methods

Participants

187 patients with medication-resistant epilepsy underwent neurosurgical procedures to implant intracranial electrodes (subdural, depth, or both) to determine epileptogenic regions. Data reported were collected at Dartmouth Medical Center (Hanover, NH), Emory University Hospital (Atlanta, Georgia), Hospital of the University of Pennsylvania (Philadelphia, PA), Mayo Clinic (Rochester, MN), Thomas Jefferson University Hospital (Philadelphia, PA), Columbia University Medical Center (New York, NY), and University of Texas Southwestern Medical Center (Dallas, TX). Prior to data collection, the research protocol was approved by the institutional review board at each hospital. Informed consent was obtained from either the participant or their guardians. Hemisphere dominance was determined by assessing handedness, Wada test (Wada and Rasmussen, 1960), or fMRI data collected during a verb generation task. Previous publications utilizing a subset of these data (88 patients) have characterized the electrophysiological signatures of successful memory encoding and retrieval (Burke et al., 2014a,b; Long et al., 2014); however, the present analyses describing the relationship of successful encoding and retrieval are novel. Deidentified raw data and analysis code used in the present analyses is available at http://memory.psych.upenn.edu/Electrophysiological_Data.

Free-recall task

Each subject performed a variant of the delayed free-recall task in which they studied a list of words with the intention to commit the items to memory. The task was performed at the bedside on a laptop, using PyEPL software (Geller et al., 2007). Analog pulses were sent to available recording channels to enable alignment of experimental events (e.g., stimulus presentation) with the recorded iEEG signal.

The recall task consisted of three distinct phases: encoding, delay, and retrieval. During encoding, lists of 12 words were presented in the native language (either English or Spanish) of the subject. Words were selected at random, without replacement, from a pool of nouns (http://memory.psych.upenn.edu/WordPools). Word presentation lasted for a duration of 1600 ms, followed by a blank inter-stimulus interval (ISI) of 750 to 1000 ms (see Fig 1a). Presentation of word lists was followed by a 20 s post-encoding delay. Subjects performed an arithmetic task during
the delay in order to disrupt memory for end-of-list items. Math problems of the form A+B+C=?? were presented to the participant, with values of A, B, and C were set to random single digit integers. Responses were made on a keypad, with presentation of additional math problems following each response (i.e., a self-paced task). After the delay, a row of asterisks, accompanied by an 800 Hz auditory tone, was presented for a duration of 300 ms to signal the start of the recall period. Subjects were instructed to recall as many words as possible from the most recent list, in any order during the 30 s recall period. Vocal responses were digitally recorded and parsed offline using Penn TotalRecall (http://memory.psych.upenn.edu/TotalRecall). Subjects performed up to 25 recall trials in a single recall session.

A subset of patients (n = 88) performed a variant (hereafter Experiment 2) of the previously described task. List presentation consisted of a total of 15 items. In addition, a green fixation cross served as a list-cue to signal an upcoming list of words. The list-cue was presented for a duration of 1600 ms, followed by the presentation of a blank screen for 800 to 1200 ms. The ISI in this variant of the task lasted from 800 to 1200 ms in duration. The recall period for this version of the task was 45 s in length.

Electrophysiological recordings and data processing

iEEG signal was recorded using subdural grids and strips (contacts spaced 10 mm apart) or depth electrodes (contacts spaced 5-10 mm apart) using recording systems at each clinical site. iEEG systems included DeltaMed XITek (Natus), Grass Telefactor, and Nihon-Kohden EEG systems. Signals were sampled at 500, 512, 1000, 1024, or 2000 Hz, depending on hardware restrictions and considerations of clinical application. Signals recorded at individual electrodes were converted to a bipolar montage by computing the difference in signal between adjacent electrode pairs on each strip, grid, and depth electrode (Burke et al., 2013). Bipolar signal was notch filtered at 60 Hz with a fourth order 2 Hz stop-band butterworth notch filter in order to remove the effects of line noise on the iEEG signal. Electrodes determined to be within the epileptogenic zone were excluded from analysis.
**Anatomical Localization**

Anatomical localization of electrode placement was accomplished using independent processing pipelines for depth and surface electrode localization. For data collected as a part of Experiment 1, post-implant CT images were coregistered with presurgical T1 and T2 weighted structural scans with Advanced Normalization Tools (Avants et al., 2008). For patients with MTL depth electrodes, hippocampal subfields and MTL cortices were automatically labeled in a pre-implant, T2-weighted MRI using the automatic segmentation of hippocampal subfields (ASHS) multi-atlas segmentation method (Yushkevich et al., 2015). Subdural electrodes were localized by reconstructing whole-brain cortical surfaces from pre-implant T1-weighted MRIs using Freesurfer (Fischl et al., 2004), and snapping electrode centroids to the cortical surface using an energy minimization algorithm (Dykstra et al., 2012). The localization for data collected in Experiment 2 differed by using FMRIB’s linear image registration tool (Jenkinson et al., 2002) for coregistration of CT and structural scans. In addition, MTL depth electrodes that were visible on CT scans were localized to either the hippocampus or PHG by neuroradiologists with expertise in MTL anatomy.

Each subject’s T1-weighted MRI was additionally registered to both an average T1 constructed from a sample of 101 patients, facilitating group-level comparisons of subdural electrodes on the cortical surface. Registration to MNI space further enabled group analysis of subdural and depth electrodes near MTL structures. Using this approach, we achieved extensive coverage across the cortical surface (Fig. 1b, left), and the MTL (Fig. 1b, right).

In order to rule out the possibility that classification of successful recall, which involves the vocalization of a verbal response, was being informed by electrophysiological biomarkers in cortical regions involved in the planning and production of a vocal response, we constructed classifiers using an anatomically restricted set of anatomical features, by excluding electrodes localized within inferior frontal gyrus (IFG) pars triangularis, IFG pars orbitalis, precentral gyrus, postcentral gyrus, paracentral gyrus, superior temporal gyrus, the bank of the superior temporal sulcus, transverse temporal gyrus, and supramarginal gyrus as defined in the Desikan-Kiliany atlas (Desikan et al., 2006).

**Spectral power**

To compute spectral power during word encoding, we applied the Morlet wavelet transform (wave number 5) to all bipolar electrode EEG signals from the onset to the offset of stimulus presentation, across 50 logarithmically spaced
frequencies from 3 to 180 Hz. Spectral power during recall was estimated from 500 ms preceding the onset of response vocalization, for correct recalls and recall errors alike. We also computed spectral power during unsuccessful periods of memory search, defined as 500 ms epochs in which the onset of any vocalization did not occur in the following 2000 ms. All events within the recall period (i.e., correct recalls and unsuccessful search) were required to be free of vocalization onsets in the preceding 2000 ms, to account for potential differences in response production across event types. Power estimates were log transformed and down sampled to 50 Hz. A buffer of 1000 ms was used to avoid edge effects on power computation. For multivariate decoding analyses, features (i.e., log-transformed power from a specific electrode and frequency band) were standardized using the mean and standard deviation estimated from all training samples, for each individual session. Test data were normalized using session means and standard deviations estimated from the training data.

By focusing on patterns of neural activity in the moments preceding individual retrieval events, we excluded recalls that were preceded by the onset of any additional vocalization or the recall period within 2000 ms. This duration was determined in order to allow estimation of low frequency power without contamination of signal from response production earlier in the recall period. While this approach decreased power in the analysis of retrieval data, it allowed for low frequency features, which have been previously demonstrated to predict successful recall (Burke et al., 2014b), to inform classification of memory states. This procedure resulted in the removal of an average of 53.8% of recall events per subject (range 6.3% to 92.7%), which had an inter-response time of 1198 ± 15.4 msec (mean ± SEM). As a result, some subjects (21 from Experiment 1, and 19 from Experiment 2) did not have a sufficient number of recall events to perform classification. For a given analysis, each subject was required to have at least 20 observations per condition. For the purposes of classification analyses, we treated unique montages (resulting from reimplants, or changes in recording electrodes) of individual patients as a single subject. This approach resulted in 202 subjects for analysis of encoding period data, and 162 subjects for analysis of retrieval period data.

Pattern classification

To identify electrophysiological biomarkers that were associated with latent memory states during the task, we trained L2-penalized logistic regression classifiers to distinguish between spectral power associated with memory success and failure. During encoding, power patterns were averaged across the 1600 ms stimulus duration at each bipolar
electrode, resulting in a set of features to train the encoding classifier. For retrieval analysis, features were constructed from the average spectral power in the 500 ms preceding response vocalization of either correct recalls or periods of unsuccessful memory search.

Classifier performance was evaluated using a cross-validation procedure. We implemented an n-fold cross-validation scheme, where n was equal to the number of completed lists. Data from one list was held out from training to test the ability of the classifier to generalize to novel data. The remaining n-1 lists were used to train the parameters of the classifier. We chose the value of the cost parameter, $C$, from one of 22 values, spaced from $10^{-4}$ to $10^6$. For each subject, $C$ was determined by selecting the value that maximized the performance of the classifier as assessed by computing the area under the receiver operating characteristic curve using list-based cross validation in the remaining sample of subjects. This approach enabled optimization of the cost parameter on an independent set of data. We trained the classifier weights by minimizing a loss function over $w$ for a set of $n$ training events:

$$\min_w \frac{w^T w}{2} + C \sum_{i=1}^{n} \log(1 + e^{-y_i w^T x_i}),$$

where $x_i$ is a set of features (i.e., the pattern of spectral power across all electrodes during encoding or retrieval), and $y_i$ is the corresponding class label for each event. Class labels during encoding were defined as 1 for recalled items, and $-1$ for forgotten items. During retrieval, correct recalls were labeled 1 and unsuccessful search periods were labeled $-1$. A bias term was included by appending a feature that was fixed to 1 for each observation. The loss function was weighted for each observation proportionally to the number of observations in each class, in order to accommodate between class imbalance. Additional classification models were constructed from both encoding and retrieval data, scaling the relative influence of each task phase by weighting retrieval observations by a scaling parameter, log spaced from $10^{-5}$ to $10^5$. We performed L2-logistic regression as implemented by the liblinear package (Fan et al., 2008).

To determine whether individual classifiers performed above chance, we first constructed the receiver operating characteristic and computed the area under the curve (AUC) across left-out samples (i.e., across the test set). We implemented a permutation test by constructing a null distribution of classifier performance by permuting the class labels within the training data. Subjects with an observed AUC in the top 95% of the null distribution were considered
significant, corresponding to a one-tailed test. To assess group performance, we used a binomial test to determine whether the proportion of subjects with significant classification would exceed the false positive rate of 0.05.

In order to rule out confounding effects of serial position and output position, we used linear mixed effects models to determine whether classifier output varied as a function of memory success (i.e., subsequent memory status or recall success). During encoding, we constructed a model with subject treated as a random effect, and memory status, list position, and the interaction between the two treated as fixed effects. Significance was determined by constructing nested models with an fixed effect term constrained to zero. After fitting the restricted model, a parametric bootstrap (n = 1000) was performed to determine whether the variance explained by the additional parameter was greater than fitting residual error with an additional parameter. We utilized a similar procedure to evaluate potential confounds in retrieval classifiers, with the fixed effects of memory status and output position.

**Identifying neural features predictive of memory success**

In order to determine which spectral features (i.e., which frequencies and regions) facilitated discrimination between successful and unsuccessful memory states, we constructed a forward model of these latent states from our linear classifiers (Haufe et al., 2014). For each subject, we computed a set of activations, \( A \), based on the learned coefficients in \( w \):

\[
A_x = \Sigma_x w \Sigma_s^{-1},
\]

(2)

where \( \Sigma_x \) is the covariance matrix of the full set of features, with each feature standardized within each experimental session, and \( \Sigma_s \) is the covariance matrix of the output of the classifier for each predicted class. For each subject and frequency, we computed an estimate of the activation at each electrode. Features were aggregated into LFA \((< 10 \text{ Hz})\) and HFA \((> 60 \text{ Hz})\) by averaging estimated activity across the frequency dimension. As a preprocessing step to the estimation of forward models, individual events were scanned for spikes and discarded if the kurtosis value for any feature exceeded a threshold of 2.5.
We projected the associated activations of each subdural electrode to all vertices on the cortical surface within 10mm of each electrode centroid. For electrodes localized within the MTL, statistical maps were constructed for each subject by assigning voxels within a 3mm radius of each electrode a value equal to its corresponding activation (overlapping regions were assigned the average of all activation values within 3mm). Each subjects activation maps was smoothed with a 4 mm gaussian filter before subsequent analysis.

Group inference was performed by nonparametric statistical testing (Maris and Oostenveld, 2007). For each test, we performed a one-sample t-test on estimated levels of activation across subjects. Resultant statistical maps were thresholded for statistical significance by constructing a null distribution obtained by random sign flipping \( n = 1000 \), and identifying positive and negative thresholds at 2.5th and 97.5th percentiles of the null distribution, ensuring a false discovery rate of 0.05.

Results

Behavioral Results

To validate patient performance, we examined characteristic behavior on the free-recall task. Subjects recalled an average of 3.26 ±0.10 (mean ± SEM) items per list, yielding an overall 24.27 ± 0.008% of items recalled. By interleaving encoding and recall with a distracting arithmetic task, we observed primacy effects in both recall initiation and overall recall rates (Fig. 2a,b). These findings are consistent with delayed free recall (Glanzer and Cunitz, 1966), and suggest that the neural mechanisms associated with memory performance reflect the encoding and retrieval of long-term memories (Strange et al., 2002). An additional behavioral marker of episodic memory function is the temporal contiguity effect (Kahana, 1996; Sederberg et al., 2010), in which subjects recall items in clusters based on their temporal proximity in the study list. Subjects recalled items with temporal organization (Fig. 2c), indicated by an average temporal factor score of 0.66 ± 0.01, which reflects significantly greater temporal clustering than chance levels of 0.5 \( t_{191} = 22.46, p < 10^{-10} \); see Polyn et al., 2009). These findings indicate normal task performance, despite relatively low levels of overall recall, consistent with impaired memory performance in epilepsy (Hermann et al., 2008).
Multiple factors contribute to performance on the recall task, including the ability to target and retrieve individual items from the previous list. When memory fails due to the inability to retrieve task-relevant information, studied items are omitted from recall. As shown in Figure 2d, the time to first recall was inversely related to the number of items yet to be recalled, and the amount of time between successive recalls increased exponentially in the recall sequences, consistent with search dynamics commonly observed in single-trial free recall (Murdock and Okada, 1970; Rohrer and Wixted, 1994). Of relevance for analysis of neural activity during memory search, inter-response times greater than 2 seconds were distributed throughout the majority of recall sequences, except early on in the recall sequence for subjects with a high rate of recall.

Memory failure during retrieval can also occur when items not studied on the previous list can be endorsed as targets, resulting in false recall. We observed false memories in the form of extra-list intrusions (ELIs), prior list intrusions (PLIs) and repetitions of previously recalled items. Recall errors were primarily ELIs, with an average rate of 0.88 ± 0.10 ELIs per list. PLIs were observed with less frequency, with an average rate of 0.57 ± 0.03 intrusions per list. The remaining errors were repetitions, which we rarely observed, with an average of 0.17 ± 0.03 of these errors made per list.

**Multivariate classification results**

In order to determine whether the same neural mechanisms support episodic memory encoding and retrieval, we constructed subject-specific classifiers that were capable of predicting latent cognitive states from patterns of spectral power across multiple electrodes and frequencies. First, we developed independent classifiers of memory success during either the encoding or retrieval of individual items. We tested the degree to which these biomarkers generalized across different phases of the task, allowing us to determine whether common biomarkers reflect successful encoding and retrieval. Finally, we constructed classification models that incorporated information from both the encoding and retrieval phases of the task, allowing us to determine which electrophysiological features were specific to either encoding or retrieval, or reflect common biomarkers of episodic memory function.
Decoding of successful memory formation and retrieval

We first constructed a classifier to identify patterns of spectral power associated with successful encoding. This classifier provides an item level estimate of the probability that subsequent remembering will occur. To relate the output of this classifier to variability in recall performance, we separated encoding events into terciles based on classifier output. By computing the probability of recall for items encoded in each of these terciles, we translated the output of the classifier to a measure of actual memory performance. We observed an $80.53 \pm 2.97\%$ difference in the proportion of items recalled in the upper, relative to the lower tercile of classifier output (Fig. 3c). The difference in recall performance was evident across all list positions, although smaller in magnitude for items encoded at the beginning of the list (Figure 3a). We observed significant classification ($p < 0.05$, permutation test) in 168 out of 202 subjects (treating changes in electrode coverage as independent subjects), with an average AUC of $0.64 \pm 0.005$. The number of subjects with significant classification performance was more than one would expect by chance ($p < 10^{-10}$, binomial test).

As the probability of recalling an item decreased across the encoding interval, we conducted a control analysis to rule out the possibility that processes that varied as a function of list position, but not memory specific processing, were primarily informing the classifier. We constructed a linear mixed effects model, predicting the evidence of each item being subsequently remembered (i.e., the exponential term in Equation 1) from both subsequent memory status and the serial position of each item. We tested whether unique variance in predicted memory states (i.e., classifier output) was associated with each factor by comparing fitness of the full model to models in which the effect of each factor was restricted to zero. We observed a significant effect of list position ($\chi^2_1 = 122.59$, $p < 0.001$ bootstrap test) and subsequent memory status ($\chi^2_1 = 188.43$, $p < 0.001$ bootstrap test). These findings confirm that the ability to identify encoding periods with a high probability of encoding success was independent of simple serial position effects that influence recall behavior.

Next, we developed classifiers to discriminate between patterns of neural activity during epochs of memory search associated with successful and unsuccessful retrieval (i.e., periods of memory search that did produce any form of recall
response). The relationship between successful retrieval and shifts in neural activity was observed across different periods of memory search (Figure 3b), with the biggest difference observed later in the recall period. Of the 162 classifiers constructed to predict retrieval success for individual subjects, we observed above chance classification ($p < 0.05$, permutation test) in 155, indicative of significant group-level performance ($p < 10^{-10}$, binomial test). Classifier performance for held out lists of items was evaluated using AUC, with average of 0.83 ± 0.01, indicating robust performance. We tested whether variability in memory states estimated by the retrieval success classifier could be predicted from the period in which an event occurred (early, middle, or late epochs defined by terciles of event onsets within the search period) and whether retrieval was successful or not. We observed significant effects for both search period ($\chi^2_1 = 14.32, p < 0.001$ bootstrap test) and retrieval status ($\chi^2_1 = 340.8, p < 0.001$ bootstrap test). These findings establish that the biomarkers of successful retrieval are informative across the duration of the search period, and are more sensitive as the retrieval period progresses. When sorting memory search epochs into terciles by classifier output (Figure 3c), subjects were 189.3 ± 5.0% more likely to recall an item when in the upper, compared to the lower tercile of classifier output.

These results suggest that memory related processing across cortical and subcortical sites can be used to reliably classify when retrieval will occur during memory search; however, such classification may be informed by neural features predictive of processes that occur subsequent to retrieval itself, including phonological retrieval and the planning of a vocal response. In order to rule out the possibility that the ability to decode successful retrieval was not primarily driven by difference in response-related activity, we constructed additional classifiers that excluded ROIs associated with vocal response production and planning (Hermes et al., 2014). Performance of classifiers constructed from this restricted feature space was observed to be significantly above chance ($p < 0.05$, permutation test) in 151 out of 162 subjects. Furthermore, we observed an average AUC of 0.79 ± 0.01, demonstrating a significant reduction from models with an unrestricted feature space ($t_{161} = -10.3, p < 10^{-18}$). These findings suggest that classification of retrieval states is not driven by response related features alone.

These findings replicate previous studies that show increased HFA, concurrent with decreased low frequency power across prefrontal, temporal, and hippocampal sites predicts the formation of episodic memories (Burke et al., 2014a; Long et al., 2014). By incorporating these signals into multivariate classifiers, we demonstrate the ability to predict
moment-to-moment variability in encoding and retrieval processes based on global brain dynamics.

**Identifying general biomarkers of memory function**

Having constructed independent classifiers that were capable of estimating the probability of memory success during either encoding or retrieval, we next sought to determine the degree to which these models reflect common or distinct processes during each phase of the free-recall task. We evaluated the degree to which the previously trained classifiers generalized across different phases of the free-recall task by evaluating classifier AUC on held out lists. As shown in Figure 4a, while decoding of memory states was less reliable when performed across phases of the task, we observed above chance classifier performance using this cross-decoding approach, providing evidence that common biomarkers of successful memory function emerge during both encoding and retrieval phases of the task. When we tested the encoding success classifier to periods of memory search associated with successful and unsuccessful memory search in held out lists of data, we observed an average AUC of 0.68 ± 0.01. Application of the retrieval success classifier to average patterns of activity during encoding yielded an average AUC of 0.59 ± 0.01.

While at first glance our cross-decoding results could be interpreted to suggest that the retrieval success model identifies biomarkers that are more specific to the memory search phase of the task than the encoding model, the differences in classifier generalization may result from a higher degree of separability in neural states in different phases of the task. We assessed the degree to which biomarkers of memory success varied as a function of classifier (i.e., encoding or retrieval success) and phase of the task using a linear mixed effects model. We observed significant effects of classifier type ($\chi^2 = 178.69, p < 0.001$ bootstrap test), task phase ($\chi^2 = 295.25, p < 0.001$ bootstrap test), and a significant interaction between the two ($\chi^2 = 177.37, p < 0.001$ bootstrap test). A post-hoc test confirmed that the decrease in classifier performance when decoding memory states across phases of the task was greater during the retrieval period ($t_{161} = 11.71, p < 10^{-22}$).

These findings suggest the presence of common biomarkers of mnemonic success during both encoding and retrieval, in addition to neural features that are specific to each phase of the task. In order to distinguish between these two
types of neural activity, we implemented classifiers that were trained on both encoding and retrieval events, and weighted the relative contribution of each phase of the task to learning. As shown in Figure 4b, identification of successful encoding states, as determined by subsequent memory, was modulated by the incorporation of neural activity associated with successful retrieval. When learning was biased towards retrieval processing (i.e., $w_{\text{enc}}/w_{\text{ret}} < 1$), we observed significant ($p < 0.05$, FDR corrected) decreases in prediction of subsequent memory, consistent with our cross-decoding results. Remarkably, performance of joint classifiers were significantly better than the encoding classifier alone at identifying optimal encoding states (Fig. 4b, red line). We next examined the ability of joint classifiers to distinguish between successful and unsuccessful memory search. We observed a significant ($p < 0.05$, FDR corrected) decrease in classifier performance compared to our retrieval classifier (Fig. 4c, blue line) when learning was biased towards patterns of neural activity present during the encoding period. These results demonstrate that the profiles of neural activation that indicate successful retrieval can be leveraged to build a more accurate description of memory states during learning, based on common changes in the local field potential.

In order to identify which regions of the brain exhibit changes in activity that reflect general biomarkers of successful memory function, we reconstructed patterns of activity that covaried with memory outcomes from joint models with the most extreme weightings (i.e., $w_{\text{enc}}/w_{\text{rec}} = 0.04$ and $w_{\text{enc}}/w_{\text{rec}} = 25.12$), and performed a conjunction analysis ($p < 0.05$ FDR corrected, permutation procedure for each independent contrast). As shown in Figure 4d, we observed common biomarkers in the form of increased HFA in inferior prefrontal, temporal, hippocampal, and parietal cortices. Decreases in low frequency activity associated with memory success were observed across multiple cortical regions, including bilateral prefrontal, temporal, and inferior parietal regions. Within the MTL, robust decreases in LFA were associated with goodness of memory in left lateralized hippocampus, and bilateral PHG.

In addition to these common features, the observed differences in classification performance across phases of the task suggest the existence of features that are specific to the encoding and retrieval of list items. We identified patterns of spectral power associated with successful memory function that were specific to either encoding or retrieval phases of the task. As depicted in Figure 5a, we observed significantly greater HFA within lateral occipital regions that was specific to successful encoding. In contrast to the typical marker of increased cortical activity supporting memory outcomes, decreased HFA within the right anterior PFC was additionally found to be a marker of successful encoding.
Decreased LFA in widespread cortical regions, as well as bilateral MTL regions, was also found to be specific to successful encoding.

[Figure 5 about here.]

Biomarkers specific to successful retrieval, predominantly increases in HFA and decreases in LFA, were observed across multiple cortical regions (Figure 5b). Of note, bilateral regions of the anterior hippocampus exhibited increased HFA that was specific to retrieval, in addition to bilateral dorsolateral prefrontal cortex, and pre- and postcentral gyri. Activation of motor cortex likely reflects movement of the throat and mouth prior to vocalization, as well as potential smearing of signal from the beginning of vocalization. Significant decreases in LFA were observed in similar cortical regions, including widespread bilateral prefrontal cortex, in addition to inferior parietal cortex and anterior temporal lobe regions.

**Discussion**

Through the multivariate classification of iEEG recorded neural activity, observed during episodic encoding and retrieval, we identified neural states that predicted moment-to-moment changes in memory function. By implementing a cross-decoding approach, we tested the correspondence between the processes that occur during encoding and retrieval, and were able to determine a high degree of overlap between biomarkers of successful encoding and recall. At a broad level, our findings suggest that successful encoding and retrieval rely upon the function of common neural substrates, a putative core episodic memory network, and that variability in the activation of this network predicts the mnemonic fate of processed information.

*Correspondence between neural processes engaged during encoding and retrieval*

Recent approaches to understanding memory function in the human brain have used multivariate classification techniques to characterize neural mechanisms involved in the formation and retrieval of episodic memories (e.g., Kuhl et al., 2012; Kuhl and Chun, 2014). While there is theoretical consensus regarding how the contents of memories are represented in the brain (Rissman and Wagner, 2012), it remains an outstanding question whether activation of common neural pathways can facilitate both episodic encoding and retrieval. Consistent with multiple neuroimaging studies that have
compared the formation and retrieval of episodic memories (Zeineh et al., 2003; Eldridge et al., 2005), we observed
activation in neural systems that were specific to either the encoding or recall of verbal information. By using a
multivariate approach to estimating successful memory processing, we demonstrate that in spite of these dissociations,
the same patterns of neural activity observed across multiple recording sites and frequencies were generally predictive
of memory function. In doing so, we have identified a putative core memory network spanning the lateral prefrontal
cortex, lateral temporal cortex, and MTL whose activation is associated with enhanced memory performance.

The common increases in HFA within this network may reflect the engagement of similar cognitive operations across
different phases of the tasks, utilizing HFA as a localizer of neuronal activity (Burke et al., 2015b). One potential
cognitive process that may account for variability in memory performance is goal-direction attention (Corbetta and
Shulman, 2002), which would modulate processing of relevant information during the memory task, including the
selection of contextually valid information during recall (Cabeza et al., 2008). As an alternative to an attention based
account, common neural activation may reflect cognitive operations supporting the maintenance and integration of
episodic content (Polyn and Kahana, 2008). Recent work examining the electrophysiological correlates of episodic
encoding (Long and Kahana, 2015) argues that increased neural activity within left prefrontal, lateral temporal, and
MTL sites reflects processing that supports the formation of episodic memories (i.e., the association of information
within a spatiotemporal context), as it predicts subsequent temporal organization of learned information. Our findings
build upon this work, and suggest that the operations supported by this network are not specific to the encoding of
memories, as they facilitate the retrieval of previously learned content.

We found that the features informing classification were consistent with recent univariate approaches to characterizing
the electrophysiological correlates of episodic encoding (Burke et al., 2014a; Greenberg et al., 2015) and retrieval
(Sederberg et al., 2007; Burke et al., 2015a) in the form of distributed networks that exhibited increases in HFA
with concomitant decreases in low frequency power that were indicative of successful memory function. Our results
also reveal differences between the neural underpinnings of encoding and retrieval, supporting theoretical models
of brain function that emphasize a division of neural resources during each operation. Specific to retrieval, we
observed increased HFA within right prefrontal cortex and the MTL, consistent with longstanding neuroimaging
findings of asymmetric hemispheric activation during episodic encoding and retrieval (Nyberg et al., 1996). While
neural processing may facilitate memory function on a specific phase of the task, such as the observed activation of ventral visual regions during encoding (see also, Burke et al., 2014a), our cross-classification findings demonstrate that similar processing states give rise to successful memory encoding and retrieval. One caveat to this correspondence results from our choice to restrict our analysis to changes in spectral power, leveraging its utility to serve as a marker of population level neural activity near the recording site. In addition to changes in levels of local neural activation, functional connectivity between neuronal populations has been shown predict the formation (Ranganath et al., 2005; Fell et al., 2008) and retrieval (Watrous et al., 2013; Kragel and Polyn, 2015) of episodic memories. As our biomarkers predominantly reflect processing within sampled neuronal populations, connectivity states that predict performance during encoding and retrieval may differ (Huijbbers et al., 2011; Duncan et al., 2014).

The models of latent memory states that we have developed can inform the development of closed-loop systems for memory enhancement. Applications of real-time classification have demonstrated the utility in the real-time detection of attentional lapses to enhance learning by manipulating task difficulty (deBettencourt et al., 2015). Attempts to enhance memory through invasive and noninvasive modulation of neural activity (for a review, see Kim et al., 2016) may prove more efficacious by accounting for the state of the brain when stimulation is applied. Recent work has demonstrated that the ability to enhance memory performance via direct brain stimulation during encoding is dependent upon the state of the brain (Ezzyat et al., 2015). Given the correspondence between electrophysiological states and proper memory function, our findings suggest that perturbation of this network during retrieval, in addition to encoding, is likely to modulate memory function.

Decoding memory states from patterns of neural activity

A challenge to the study of human memory is to determine the processes that are responsible for both the transformation of experience into lasting memories as well as the processes involved in retrieving encoded information. Neuroscientific endeavors to better elucidate these processes often rely on contrasting markers of neural function based upon the success of an attempt to remember. For example, the subsequent memory paradigm (Paller and Wagner, 2002) has been utilized to reveal the contributions of prefrontal, MTL, and parietal cortex to the process of memory formation in fMRI (Wagner et al., 1998; Davachi et al., 2001; Kim, 2011) and electrophysiological studies of human memory (Sederberg et al., 2007; Long et al., 2014). While these approaches have established a foundation with which to
understand the neural processes that underpin human memory, the correspondence between neural states and general memory performance is often overlooked (c.f., Hariri et al., 2003). In contrast, our classification models provide estimates of the probability of memory succeeding or failing based on the observed electrophysiologic state of the brain, effectively providing an estimate of latent cognitive states related to memory processing. This approach enabled us to determine the contributions of processing during encoding and retrieval periods to memory performance.

We observed the greatest correspondence between patterns of neural activity and behavior during successful retrieval, as compared to successful encoding. While it is tempting to interpret these results as indicating that neural processing during retrieval, rather than encoding, plays a greater role in determining the ability to remember (for a theoretical perspective on the role of retrieval processes in remembering, see Tulving, 1974), we do not believe this to be the case. If one assumes our models track the engagement of neural mechanisms supporting multiple cognitive processes requisite for task performance (e.g., the maintenance of memory representations during encoding and retrieval), then classifier performance by necessity reflects the degree to which these processes are differentially engaged across observations associated with successful or unsuccessful memory function. In this light, improved classification of successful retrieval results from a greater correspondence between the outcome of memory search (i.e., successful recall or failure) and underlying neurophysiology. The inability to achieve comparable classifier performance at predicting subsequent memory results from items that were forgotten despite appropriate neural processing at the time of encoding. During the free-recall task, wherein subjects are free to determine which cues they use to probe memory, memory can fail due to ineffective self-generated retrieval cues. As it is challenging to determine the retrieval cues used by a subject during a specific recall period (c.f., Polyn et al., 2005), future work should examine the relative contributions of encoding processes when retrieval cues are under experimental control. This could rule out variability in memory performance due to inefficient use of retrieval cues during recall, a phenomenon observed in patients with prefrontal damage (Stuss et al., 1994).

An alternative interpretation of our findings is that additional episodic encoding occurs in the moments leading up to recall. Indeed, the act of retrieval is known to enhance learning, with activation of inferior prefrontal and inferior temporal regions (Buckner et al., 2001) as well as structures within the MTL (Stark and Okado, 2003) predicting the degree to which novel recognition probes are learned during retrieval tasks. We believe it is unlikely that the present
results solely reflect incidental encoding processes engaged during free recall. It has been demonstrated that processing within the prefrontal cortex and MTL that predicts subsequent remembering is sensitive to the novelty of learned information (Kirchhoff et al., 2000), with MTL activation attenuating in response to the repetition of well encoded stimuli (Turk-Browne et al., 2006). As retrieval of encoded information results from the reactivation of previously formed item representations, processes associated with the encoding of novel stimuli into long-term memory are unlikely to be engaged during memory search (Kragel and Polyn, 2016).

Conclusion

The processes involved in encoding an event into a lasting memory and remembering that event at a later point in time rely upon inherently distinct neural mechanisms. We demonstrated that multivariate models of memory success can generalize across the encoding and recall of verbal information. These findings suggest that in the presence of neural processes that are specific to the encoding and recall of episodic memories, engagement of a putative core memory system generally shapes the ability to remember the past.
References


Huijbers W, Pennartz CM, Cabeza R, Daselaar SM. 2011. The hippocampus is coupled with the default network during memory retrieval but not during memory encoding. PloS one. 6:e17463.


Figure 1. Recall task and electrode coverage. a. Experimental paradigm. On each trial, patients studied a list of words, performed a self-paced arithmetic task, and finally recalled the items studied on the most recent list in any order. b. Electrode coverage. Left, the number of subjects with an electrode within 10mm of each vertex of the average cortical surface. Right, cross sections along the longitudinal axis of the MTL showing the number of subjects with electrode coverage localized within hippocampus (Hipp) or parahippocampal gyrus (PHG), extending 3mm from electrode centroids. L, left; R, right.
Figure 2. Behavioral results. a. The overall probability that items from each list position were recalled. b. The probability that an item from a given list position was the first item recalled, as a function of serial position. c. The probability that a recall transition will come from a temporally proximal list position, given it is available for recall. d. Inter-response time (measured from onset to onset of response vocalization) as a function of recall sequence length. Error bars represent standard error of the mean.
Figure 3. **Decoding successful memory formation and retrieval.** a. Probability of encoding success as a function of list position. The top panel depicts the probability of recalling an item as a function of serial position, for each tercile of classifier output. Below, the difference in percent recall change (relative to average percent recall) between the upper and lower tercile of classifier output. b. Impact of retrieval success biomarkers increases across the memory search period. The top panel shows the average proportion of recalls made in epochs decoded across the recall period, sorted into terciles based upon classifier output. Below, the difference in the proportion of recalls made relative to the mean between upper and lower terciles of classifier output. c. Overall change in memory performance during encoding (ENC) and retrieval (RET) between high and low biomarker states. Error bars represent standard error of the mean.
Figure 4. **Common biomarkers of successful memory function.**

- **a.** Classifier performance as a function of task phase. Average area under the curve (AUC) for classifiers trained to detect successful encoding and retrieval processing. Error bars reflect standard error of the mean, across subjects.

- **b.** Predicting successful encoding using a joint classifier. AUC for joint classifiers applied to encoding period data, with significant ($p < 0.05$, FDR corrected) increases and decreases in performance, compared to the encoding classifier (dashed line) are shown in red and blue, respectively. Shaded area depicts standard error of the mean.

- **c.** Predicting successful retrieval using a joint classifier. AUC for joint classifiers applied to retrieval period observations, with significant decreases in performance compared to the retrieval classifier alone (dashed line) are depicted in blue ($p < 0.05$, FDR corrected).

- **d.** Increases in high frequency activity (HFA) and decreases in low frequency activity (LFA) are common biomarkers of memory encoding and retrieval. Statistical maps of the cortical surface (left) and MTL subregions (right) indicating significant increases (red) and decreases (blue) in spectral power associated with successful memory function ($p < 0.05$, FDR corrected).
Figure 5. **Encoding and retrieval specific biomarkers.**  

**a.** Changes in activity that are specific to successful memory encoding. Regions with activation that significantly differs between subsequently remembered and forgotten items, exclusively masking out any regions that exhibit retrieval related effects are shown. 

**b.** Cortical and MTL regions showing significant differences in activity, specific to episodic retrieval. Significant ($p < 0.05$, FDR corrected) increases in high frequency activity (HFA) and decreases in low frequency activity (LFA) are depicted on the cortical surface (left) and MTL subregions (right) are shown in red and blue, respectively.