MODELS AND MECHANISMS FOR INTEGRATING CORTICAL FEATURE SPACES

by

Joseph Daniel Monaco

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

in the Graduate School of Arts and Sciences

Columbia University

2009

ABSTRACT

MODELS AND MECHANISMS FOR INTEGRATING CORTICAL FEATURE SPACES

Joseph Daniel Monaco

The medial temporal lobe memory system is the seat of explicit memory in mammals, including both recognition and episodic memory. Recognition is the sense that an object, a scene, or some other context has been experienced previously. Episodic memory is the automatic and long-term storage of life's experiences, each separate and distinct. These very different forms of memory are served by different subsystems in the medial temporal lobe and hippocampal formation. The difference in computational function may result from differences in the structure of the inputs that drive their respective networks. We examine models and mechanisms participating in both of these memory subsystems. Particularly, we consider how the high-order feature spaces that they integrate participate in the underlying computation. The primate perirhinal cortex computes a familiarity signal for high-order objective features, an input space with structured correlations. We present a minimal model of this familiarity signal and show that word-frequency information is encoded in the space of semantic similarity. The hippocampus in rodents is critical to their ability to navigate through the

world and learn their spatial surroundings, but it also provides a model for episodic memory. We present a mechanism that provides rapid sparsification of the spatial inputs to hippocampus and produces informative spatial representations. This mechanism may enable fast contextual learning in unfamiliar environments and, simultaneously, provide the distinct basis of activity necessary for encoding episodic memory.

Contents

List of Figures

ACKNOWLEDGMENTS

I want to thank my graduate advisor Larry Abbott, who maintained his faith in me when I needed it and helped me learn to keep things simple. I would not have been in a place to do this work without my undergraduate advisor Chip Levy, who taught me the importance of clear writing and knowing what you're talking about. I have learned a lot from my collaborators on these and other projects, including Mike Kahana and Isabel Muzzio. I greatly appreciate my thesis advisory committee members for asking the right questions: Steve Siegelbaum, Josh Gordon and Ken Miller. I am indebted to the outside examiners who took part in the defense committee: Aurel Lazar and Mike Hasselmo. I owe great thanks to the administrators who kept things (and me) running smoothly: Monique Dols, Cecil Oberbeck and Alla Kerzhner.

I want to thank Pablo Jercog for very helpful conversations that kept my thinking on track. I'm grateful for my former colleagues Tim Vogels, Josh Jacobs and Patryk Laurent, who have become good friends; and current colleagues George Jun Zhao, Sean Luo, Xaq Pitkow, Michael Vidne, Clay Lacefield and others who made the trip worthwhile.

For my parents and step-parents, I'm sorry it took so long, but thanks for waiting; enormous thanks for your support, which has taken many forms. And for my brother Ben, thanks for letting me play the latest video games once in a while.

For Grace, who continues to surprise me.

Chapter 1

Small Worlds and Random Graphs for Recognition and Recall

The sight of the little madeleine had recalled nothing to my mind before I tasted it; perhaps because I had so often seen such things in the interval, without tasting them, on the trays in pastry-cooks' windows, that their image had dissociated itself from those Combray days to take its place among others more recent; perhaps because of those memories, so long abandoned and put out of mind, nothing now survived, everything was scattered; the forms of things, including that of the little scallop-shell of pastry, so richly sensual under its severe, religious folds, were either obliterated or had been so long dormant as to have lost the power of expansion which would have allowed them to resume their place in my consciousness.

—Marcel Proust, *À la recherche du temps perdu*

Summary

We introduce the broad themes and concepts investigated in later chapters from the perspective of mnemonic function. First, we describe the medial temporal lobe and its support for both recognition and episodic memory. We discuss the view supporting parallel streams of cortical input to the hippocampal formation in terms of the differing encoding/retrieval requirements for these types of memory. Next, we review familiarity discrimination in the perirhinal cortex of primates and discuss the structure of the feature spaces that it integrates. Finally, we review the entorhinal–hippocampal system of spatial representation in rodents as a model for understanding episodic memory.

1.1 Introduction

It has been widely recognized that the various subregions of the hippocampal formation (HF) and the surrounding parahippocampal areas (PHA) are the seat of explicit, or declarative, memory. These medial temporal lobe (MTL) structures enable diverse capabilities relating to the long-term storage of experiential and semantic knowledge (Squire and Zola, 1998; Squire et al., 2004). This functional diversity is reflected in the complexity of anatomy, connectivity and integrative processing of high-order neocortical inputs evident among the subregions of the parahippocampal–hippocampal network (Figure 1.1; van Strien et al., 2009).

The study of the famous patient H. M., now known to have been Henry Gustav Molaison (1926–2008), demonstrated the functional and anatomical separability of the declarative memory system from that of implicit, procedural and motor learning (Scoville and Milner, 1957; Carey, 2008). His extensive bilateral lesions included the HF and most of the MTL, resulting in the inability to form new

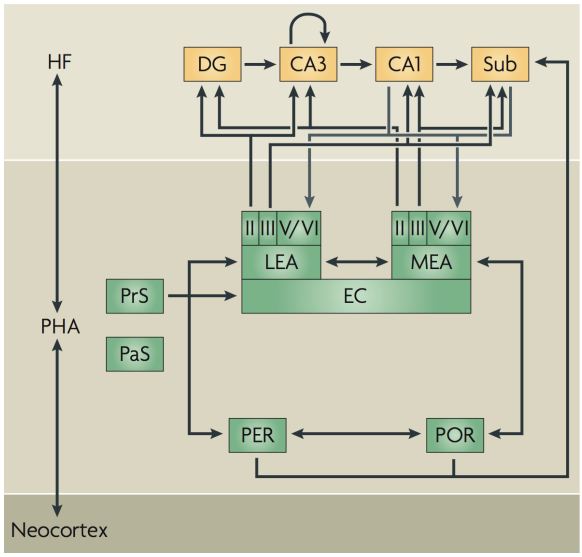

Figure 1.1 Overview of the main connections among the hippocampal subnetworks and adjacent parahippocampal areas. This diagram illustrates the conventional view that neocortex feeds multimodal input to the PHA cortical areas, which are reciprocally connected to the HF via the interface between entorhinal cortex (EC) and the "trisynaptic loop" of hippocampus consisting of DG, CA3, CA1, and subiculum (Sub). Parallel processing paths are emphasized: perirhinal (PER) to the lateral entorhinal area (LEA); postrhinal (POR) to the medial entorhinal area (MEA). Other parahippocampal areas shown are the presubiculum (PrS) and parasubiculum (PaS). Adapted from van Strien et al. (2009).

episodic memories; however, the relatively minimal damage to his lateral temporal cortex may have preserved his semantic memory (Schmolck et al., 2002; Levy et al., 2004). This illustrates a further dissociation within the MTL memory system: that between episodic memory, for one-time events, and semantic and recognition memory, for facts, objects, and familiarity judgments relating to stimuli that may be experienced repeatedly over a lifetime. The HF and especially the subregions CA3 and CA1 of hippocampus are likely to be involved in the item–context associations necessary for efficiently encoding and storing single episodes (Eichenbaum, 1999; Knierim et al., 2006; Colgin et al., 2008). Recognition memory has been extensively characterized as a dual mechanism consisting of familiarity discrimination

and recollection (Yonelinas, 2002; Rugg and Yonelinas, 2003), with the latter involving contextual retrieval whereas familiarity refers to the sense or judgment of prior occurrence. Neural responses consistent with recognition have been observed throughout the HF and adjacent cortical areas in MTL (e.g., Fahy et al., 1993), but perirhinal cortex is undisputedly central to high-order, object-specific recognition (Suzuki, 1996; Murray and Bussey, 1999; Murray et al., 2007; Squire et al., 2007). Hippocampus also participates in recognition, though responses there tend to more abstractly signal novelty and usually in conjunction with other high-level features such as task requirements (Levy et al., 2004).

We consider these forms of memory, familiarity-based recognition and episodic memory, in the context of the structure–function relationships with their respective high-order cortical inputs. Both mechanisms ultimately integrate and associate high-level feature spaces, whether semantic, objective, or spatial, that are presented to their respective mnemonic networks as experiences occur. According to the conventional view, afferents from the top of hierarchically-organized sensory pathways in neocortex project to the PHA subnetworks, and these then reciprocally innervate the hippocampal subregions largely via entorhinal cortex (Canto et al., 2008; van Strien et al., 2009). This view further posits parallel streams of input, in which the perirhinal and postrhinal cortices are points of convergence for neocortical afferents that project, respectively, to lateral and medial entorhinal cortex (Figure 1.1). In rat hippocampus, this entails a dissociation between spatial and non-spatial inputs (Hargreaves et al., 2005) that may form the basis for separate object and space pathways underlying a putative rodent analog of episodic memory (Eichenbaum, 1999; Knierim et al., 2006). Given this, and somewhat simplistically, we can say that familiarity-based recognition and the context-dependent encoding of episodic memory operate both at different levels of abstraction and depend

on different processing streams. That is, object familiarity operates based on direct neocortical inputs in the lateral (object) stream, while contextual association in hippocampus relies on higher-level entorhinal inputs from the medial (spatial) stream (Figure 1.1). Familiarity discrimination and episodic memory also have vastly different encoding and retrieval requirements, indicating that differences in the structure of their respective input spaces may reflect different computational constraints.

Efficient memory encoding to avoid interference and maximize capacity requires pattern separation. Pattern separation refers to any process that transforms similar patterns of activity into less similar, or perhaps orthogonal, output patterns. This prevents newly encoded activity patterns from interfering with those of previously stored memories; however, it also works against pattern completion, in which an associated cue or partial stimulus can induce item retrieval. With perfect pattern separation, there would be no relationship between the stimulus that induced the encoding of a memory and the activity pattern in which it was stored. This would make retrieval difficult or impossible without a specialized decoder mechanism. Such a role has been proposed for CA1 in hippocampus (Levy, 1989; O'Reilly and McClelland, 1994; Hasselmo, 1995). Although decoder models based on strong entorhinal input to CA1 have been rejected (Levy et al., 1995), cholinergic modulation of the relative strengths of its Schaffer collateral and entorhinal inputs may in fact support a decoder role for CA1 (Hasselmo and Schnell, 1994). Generally, pattern separation and completion should be balanced to minimize inefficiencies in encoding and retrieval. O'Reilly and McClelland (1994) suggested, using a "Hebb–Marr" model of the entorhinal projection to dentate gyrus (DG) and CA3, that hippocampus is structurally well-balanced in this respect. Indeed, the dual inputs to CA3, strong mossy-fiber synapses from DG and dispersed but modifi-

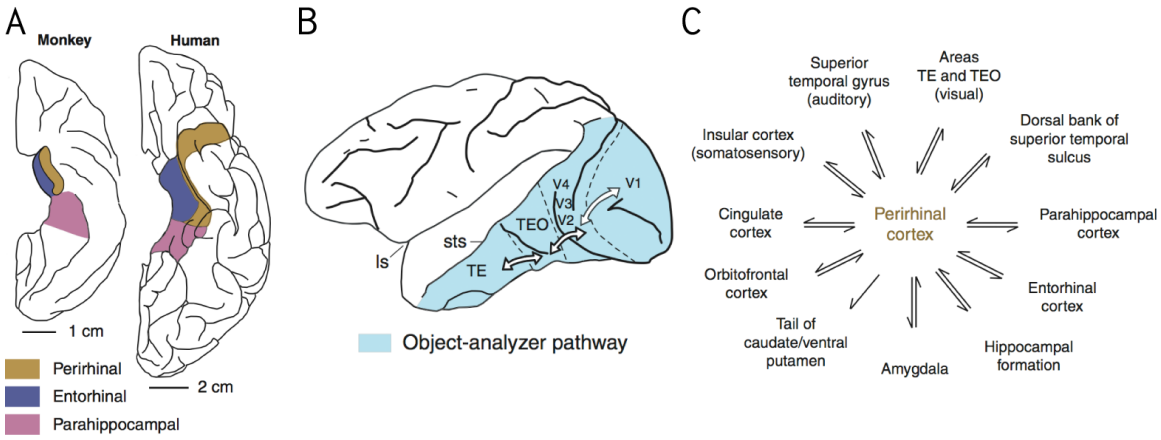

Figure 1.2 Ventral pathway carrying visual information to the temporal pole, where it converges with other sensory and polymodal pathways in perirhinal cortex. A. Diagrams of perirhinal cortex location in monkey and human (left hemisphere, ventral view, top is rostral). B. Diagram of the object-analyzer pathway in monkey, which carries information from the first-order visual cortex (V1) caudally through a sequence of higher-order cortical fields to area TE (inferotemporal cortex) rostrally. Perirhinal cortex is located medially (not shown), opposite area TE. C. Significant diversity and convergence of polymodal inputs to perirhinal cortex, making it the first field in the visual pathway to integrate information from other sensory pathways. Adapted from Murray et al. (2007).

able perforant-path synapses from entorhinal cortex (Figure 1.4c), may be uniquely suitable to minimize this encoding/retrieval trade-off (Treves and Rolls, 1992). Below, we briefly review and compare two specific mnemonic systems: familiarity discrimination in primate perirhinal cortex and spatial representations in the rodent entorhinal–hippocampal network. We discuss the structure of the respective cortical inputs with consideration for their differing separation–completion requirements.

1.2 Feature correlations and familiarity

First, we consider familiarity discrimination in perirhinal (PER) cortex of human and non-human primates. PER is located on the ventromedial aspect of the temporal lobe (Figure 1.2a), consisting of Brodmann areas 35 and 36. It extends anteriorly to the medial portion of the temporal pole (Suzuki, 1996) and far enough laterally to border area TE in monkeys and inferotemporal (IT) cortex in humans (Murray and Bussey, 1999). Indeed, PER integrates high-order visual information from the unimodal object-identification pathway (Figure 1.2b), which consists of a series of visual cortical subfields starting caudally with V1 and progressing to areas TEO and TE in monkey (Murray et al., 2007). Though the most prominent, the visual pathway is not the only cortical projection to converge in this part of the temporal lobe. PER receives largely reciprocal connections from somatosensory and auditory association areas as well as polymodal inputs such as orbitofrontal cortex (Figure 1.2c; Suzuki, 1996; Murray et al., 2007). Medially, PER shares a substantial border with the lateral aspect of entorhinal cortex (LEC) and regions within the PHA (Figure 1.2a), giving rise to strong reciprocal connectivity between PER and LEC (Figure 1.1). Thus, the two primary characteristics of perirhinal connectivity are its extreme convergence of unimodal and polymodal cortical inputs, with an emphasis on visual-object information, and its major projection to the HF through the lateral stream (Suzuki, 1996; Insausti and Amaral, 2008). That is, perirhinal cortex funnels high-bandwidth, high-order objective and experiential information into the hippocampal object-processing stream.

Neural responses indicating stimulus familiarity have been found throughout the anterior inferotemporal lobe and MTL structures such as perirhinal and entorhinal cortex (for review, see Brown and Xiang, 1998). Recordings from mon-

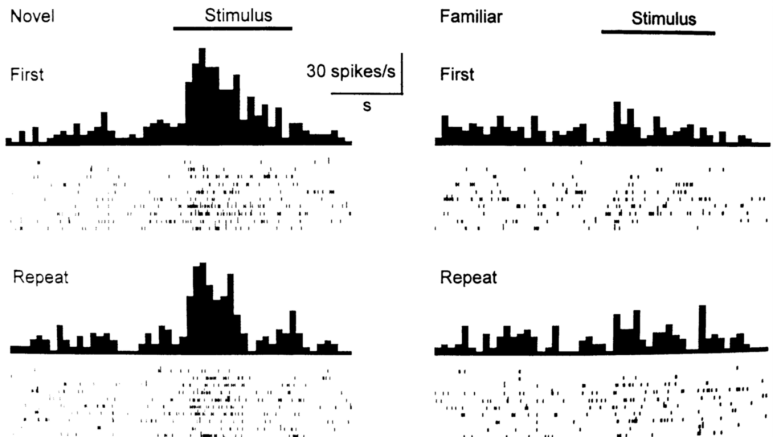

Figure 1.3 Familiarity neurons in anterior inferotemporal cortex respond with decreased firing to familiar stimuli. Area TE neurons classified as "familiarity" units, which are found throughout MTL including perirhinal cortex, have decreased responses to familiar stimuli. Within-trial repetitions of novel stimuli do not elicit this stimulus-specific decrease; the response develops 4–8 minutes after initial presentation and is still evident after 24 hours. Adapted from Xiang and Brown (1998).

keys performing visual recognition tasks found a diversity of neurons with responses indicating recency, familiarity or novelty (Miller et al., 1991; Fahy et al., 1993; Sobotka and Ringo, 1993; Xiang and Brown, 1998). The familiarity response, a putative signal for familiarity discrimination, consists of a significant decrease in firing rate upon presentation of a previously experienced stimulus (Figure 1.3). Notably, an experiment involving hemispheric transfer by transecting the corpus callosum dissociated this stimulus-specific "repetitive response suppression" from recognition performance (Sobotka and Ringo, 1996), indicating the response may not be a neural correlate of behavior. However, more recent evidence from animal lesion, electrophysiology, and even human functional imaging studies increasingly suggests both a dissociation between familiarity and recollection and that the critical neural substrate of stimulus-specific familiarity discrimination in MTL is indeed the PER response decrease (for review, see Aggleton and Brown, 2006).

For a familiarity neuron in MTL, the repeated presentation of a stimulus does

not elicit the decreased response immediately. It takes several minutes for the response to develop, after which it is evident at least 24 hours later (Xiang and Brown, 1998). This indicates that a slow, plastic process of synaptic modification may produce the familiarity response. It was initially unclear whether the site of this synaptic change was perirhinal cortex or its immediate afferent area TE: both areas produce the familiarity response and are necessary for performance on delayed-match-to-sample recognition tasks. These observations for area TE may just reflect the feedback connections from PER and the necessity of area-TE input for eliciting the PER response (Brown and Xiang, 1998). However, shorter response latencies for familiarity neurons in area TE than in PER suggest that inferotemporal activity contributes to the familiarity response, and subsequent synaptic plasticity, in PER (Xiang and Brown, 1998). Indeed, a novel form of activity-dependent long-term depression (LTD) in rat perirhinal cortex may provide a mechanism for the plastic change underlying the firing rate decrease (Cho et al., 2000).

Computationally, the familiarity discrimination process in perirhinal cortex must transform single-trial presentations of particular high-order feature vectors into a long-term learning signal that attenuates future network responses to subsequent presentations of those features. Consider the feature space integrated by a familiarity network in perirhinal cortex. It has been suggested that sensory neocortex must build stable representations of objects and modalities for association areas to have meaningful sensory information (O'Reilly and McClelland, 1994). This entails not pattern separation but the clustering of similar patterns. Indeed, high-level perceptual categorization is well-known in IT cortex (Miller et al., 2003); further, the semantic clustering of perceptual stimuli appears to be both fine-grained and hierarchically organized (Kiani et al., 2007). All of which means that semantically similar visual or objective stimuli are likely to elicit similar population responses

in monkey area TE and human IT cortex. A recollective process would require pattern separation on these correlated features to avoid interference during retrieval, but retrieval for a pure familiarity discrimination process requires just a scalar output commensurate with the relative familiarity or novelty of the stimulus. Since the context of the stimulus is irrelevant to the output, a familiarity discrimination process only needs to assess the likelihood that some combination of features was presented previously. Indeed, models of perirhinal familiarity discrimination have shown that if recollection is separate from the familiarity network, then familiarity-based recognition is very fast and has a very high capacity (Sohal and Hasselmo, 2000; Bogacz et al., 2001b; Norman and O'Reilly, 2003). Interestingly, Bogacz and Brown (2003) show that a network with anti-Hebbian synaptic modification is able to learn to ignore shared features so that it is only responsive to novel features. This enables much higher capacity than models that rely on Hebbian association to learn stimulus representations, but may actually lead to too little interference to explain known interference effects in recognition (Norman and O'Reilly, 2003).

Finally, the PER familiarity mechanism may be critically involved in lexical-semantic recognition memory: this is the memory of words and their semantic associations. MRI and functional imaging in humans demonstrate that the degree of atrophy within the anterior temporal lobe and perirhinal cortex in particular correlates with the severity of semantic dementia in patients (Hodges and Patterson, 2007). It is possible that the apparent anatomical specificity of semantic dementia may reflect correlations with more generalized neurological deterioration (Schmolck et al., 2002; Levy et al., 2004). Nonetheless, semantic dementia is characterized by deficits of lexical expressiveness, receptive vocabulary, general semantic knowledge of objects, and the ablility to define low-frequency words. Similar to perceptual object categorization, the structure of lexical feature spaces

may be highly clustered. To verify this, Steyvers and Tenenbaum (2005) performed clustering analyses on several empirically determined semantic spaces as well as a simple model of semantic growth. They found that both the empirical spaces and model semantic networks exhibited small-world statistics, characterized by short average path lengths and a high degree of clustering. To assess the effects of the detailed structure of these spaces in recognition, we studied a minimal network model of familiarity discrimination using an empirical word-association space as input (Chapter 2; Monaco et al., 2007). We found that input correlations decreased recognition capacity overall but, notably, produced qualitatively correct word-frequency effects.

1.3 Space, orthogonality and the recall of experience

Next, we consider the system of spatial representation in the medial entorhinal cortex (MEC) and the perforant path structures in hippocampus, dentate gyrus (DG) and subregion CA3. The entorhinal cortex (Figure 1.4a–b) and the perforant path (PP) projections (Figure 1.4c) to DG have been studied extensively from the beginning of neuroanatomical study (Ramón y Cajal, 1901) to the era of three-dimensional computer reconstruction techniques (Dolorfo and Amaral, 1998). Famously, Ramón y Cajal noted the critical importance of the entorhinal structure to hippocampal function by asserting that were it a visual area, so should be the hippocampus. The entorhinal cortex critically relays neocortical and PHA-originated projections through the PP to the HF. The major portion of the PP consists of axons from entorhinal layer II cells, though other layers contribute, which innervate the entire transverse axis of DG and send collaterals directly to stratum lacunosum-moleculare (s.lm.) of CA3 (Witter, 2007; Canto et al., 2008). The primary division

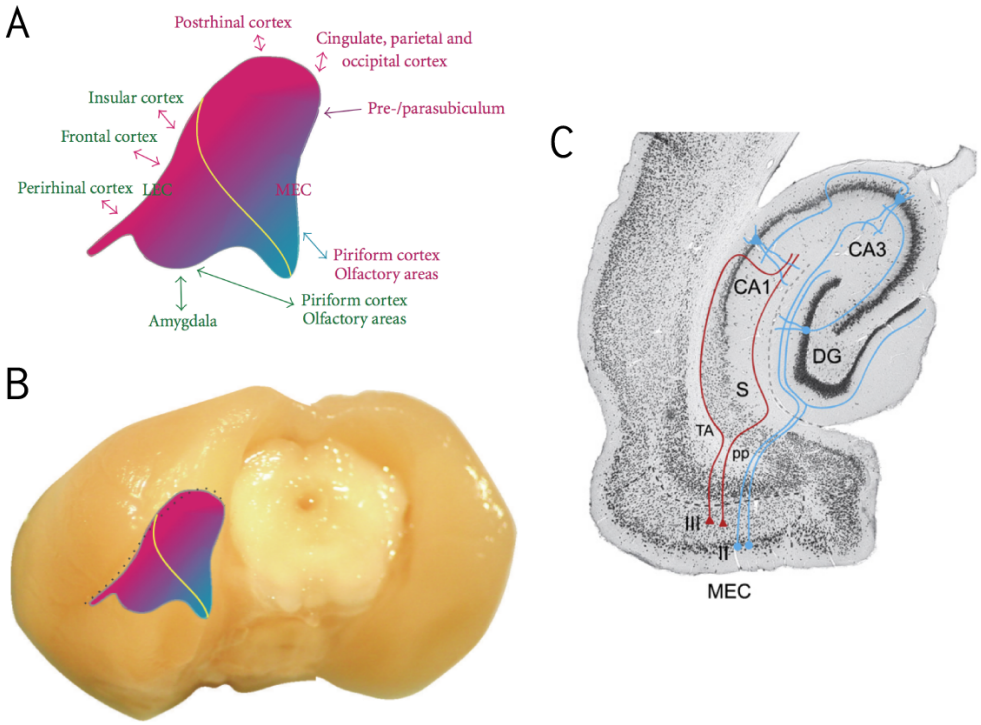

Figure 1.4 Differential connectivity of medial and lateral entorhinal cortex in the rat, with projections to the hippocampal subfields. A. Surface map of entorhinal cortex showing medial–lateral division (yellow line; the dorsolateral–ventromedial hippocampal projection gradient is colored magenta–blue). B. Entorhinal map superposed on posteriolateral view of rat brain. C. Photomicrograph of horizontal section, with schematic illustration of the perforant path (blue) and temporoammonic tract (red) projections to hippocampus. Adapted from Canto et al. (2008) (A and B) and Brun et al. (2008) (C).

in the entorhinal projection is medial–lateral: LEC projects to the outer third of the dentate molecular layer and superficial s.lm. of CA3, whereas MEC projects to the middle third of dentate molecular and deep s.lm. of CA3 (Dolorfo and Amaral, 1998; Witter, 2007; van Strien et al., 2009). This division and differential pattern of hippocampal innervation at the PHA–HF interface supports the view that parallel processing streams become associated in hippocampus (Figure 1.1). The other major topographic feature of this projection is the dorsolateral–ventromedial gradient (Figure 1.4a–b, magenta–blue), which corresponds to innervation along the

hippocampal septal–temporal axis (Dolorfo and Amaral, 1998). Thus, the septal halves of DG and CA3, where spatial activity is most prominent, receive entorhinal input primarily from a dorsolateral band across LEC and MEC.

The highly localized spatial activity of place cells in rat hippocampus is one of the most well-known and extensively studied neural correlates of behavior in neuroscience (O'Keefe and Dostrovsky, 1971; O'Keefe, 1976). The place fields of populations of place units form coherent maps of the environment (O'Keefe and Nadel, 1978) that support spatial navigation and location-dependent learning and memory. Spatial responses in DG typically consist of multiple firing fields, whereas those of the ammonic subregions CA3 and CA1 are more robustly characterized by a strong, single place field (see, e.g., Leutgeb et al., 2007). Place activity is very sparse, with typically 50–70% of the place population silent within a given environment; but it is highly informative, as only a small number of active place units are needed to achieve high-precision predictions of an animal's location (Wilson and McNaughton, 1993).

It is instructive to consider the dynamics of onset for these spatial representations. In a familiar environment, a previous representation may be autoassociatively retrieved (Marr, 1971) based on salient cues or internal states. Preliminary evidence suggests that this retrieval process demonstrates attractor dynamics and can occur rapidly on the timescale of a single theta cycle (Jezek et al., 2008). However, in a novel environment there may not be sufficient similarity with any previous representations to induce retrieval. The dynamics of spatial map formation in novel environments are therefore crucial to the encoding of spatial memory. Hill (1978) first showed that spatial activity is largely evident on an animal's first pass through a new environment, though some place cells can take several minutes to develop place activity. Recording from CA1, Wilson and McNaughton

(1993) showed that this initial ensemble activity is an unreliable predictor of location, but that it stabilizes on the timescale of tens of minutes. They also found a corresponding suppression of CA1 interneurons, later verified by Nitz and McNaughton (2004) who also found a simultaneous enhancement of interneuronal activity in DG. Frank et al. (2004) used an adaptive estimation algorithm to assess the instantaneous structure of CA1 place fields as they developed in a novel environment. They showed a diversity of substantial field changes across time and that place fields tended to stabilize with 5–6 total minutes of experience in the environment. Subsequently, a series of studies demonstrated that the active cell assemblies recruited in new environments were independently sampled and largely non-overlapping with those of familiar environments (Leutgeb et al., 2004; Vazdarjanova and Guzowski, 2004; Leutgeb et al., 2005). This results in an "orthogonalization" of spatial representations known as global remapping, which we review further in Section 4.1.1. Finally, substantial amounts of evidence strongly suggest that spatial learning across the hippocampal subregions depends on some form of novelty-evoked, NMDA-dependent plasticity (e.g., Croll et al., 1992; Nakazawa et al., 2003; Monaco and Levy, 2003; Li et al., 2003; Haberman et al., 2008). Some recent studies, further, suggest a dissociation (Leutgeb et al., 2007; Karlsson and Frank, 2008): that the initial spatial selectivity of both DG and CA1 activity seems to depend on such plasticity, while initial CA3 activity in a novel environment does not. Thus, we explore a model constrained by hard-wired network dynamics as a mechanism for the initial formation of spatial representations (Chapter 3).

The discovery of grid-like spatial firing of some MEC layer II cells (Hafting et al., 2005), one synapse upstream from DG and CA3, has revolutionized our understanding of spatial representations in the PHA–HF network. The cells were found in the dorsocaudal aspect of MEC in a dorsoventral strip starting at the rhi-

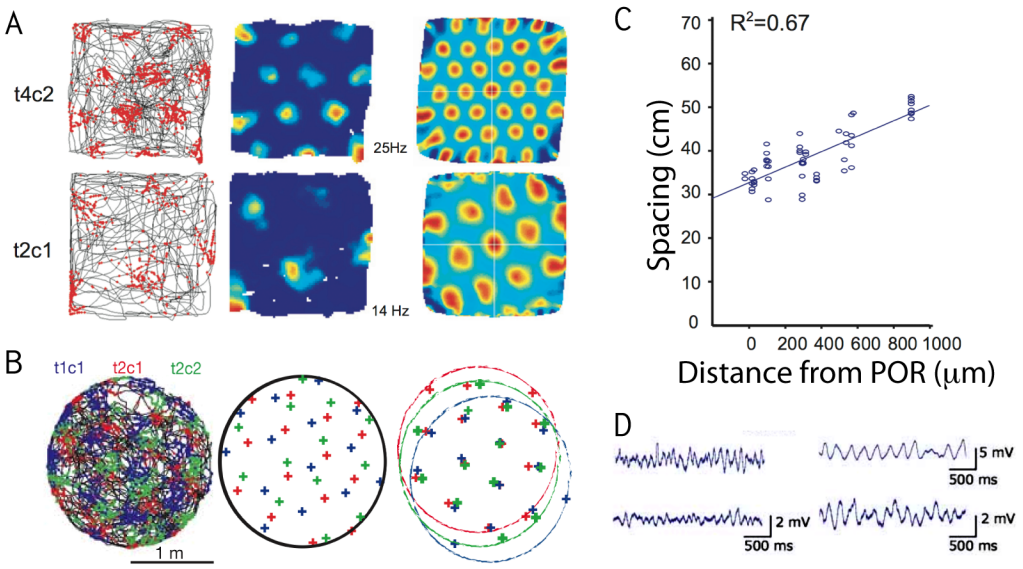

Figure 1.5 Characteristics of grid-cell responses along the spatial frequency gradient. A. MEC grid responses from dorsal (top) and ventral (bottom) aspects of dorsocaudal MEC. Trajectory with overlaid spikes are shown (left) with smoothed ratemaps (middle) and autocorrelograms (right). B. Colocalized ensembles of grid cells have diverse spatial phases, covering the environment with a few neighboring cells. C. Spacing between firing fields increases along the dorsoventral gradient (measured as distance from postrhinal (POR) border). D. Intrinsic membrane oscillation frequencies follow a similar gradient from dorsal (left) to ventral (right). Adapted from Hafting et al. (2005) (A–C) and Giocomo et al. (2007) (D).

nal sulcus (Fyhn et al., 2004), which forms the entorhinal border with postrhinal (POR) cortex (Figure 1.4a). As discussed above, both dentate granule and CA3 pyramidal cells receive a significant proportion of their afferent input from this layer of cells (Witter and Moser, 2006; Witter, 2007). Given the ordered sequence of hippocampal subregions and deep-to-superficial laminar recurrence in entorhinal cortex (Figures 1.1 and 1.4c; Canto et al., 2008), the flow of spatial information in this system follows a complicated but relatively closed-loop structure. The MEC grid-cell response consists of periodic firing fields at the vertices of a hexagonal grid that tessellates the environment (Figure 1.5a). The primary metrics of the grids are the spatial phase, measured as the offset from the center of the envi-

ronment of the nearest field; the orientation, measured as the grid's rotation from some reference angle; and the spacing between fields, which determines the spatial frequency of a grid. There is some organization to these metrics. Colocalized ensembles of grids share the same orientation and spacing. However, their spatial phases are unrelated, so that the spatial responses of a small number of neighboring grid cells are sufficient to cover the environment (Figure 1.5b). This randomness of spatial phases enables the system of grids to be characterized as part of a path-integration mechanism for navigation (McNaughton et al., 2006). The other major topographic organization is that spatial frequency decreases along a dorsoventral gradient from the POR border (Figure 1.5c). So, grid cells found dorsally (Figure 1.5a, top) have smaller grid spacing relative to grid cells found ventrally (Figure 1.5a, bottom). Modeling studies have suggested that grid responses emerge from interference between subthreshold membrane oscillations and the entorhinal theta rhythm (Burgess et al., 2007; Hasselmo et al., 2007). Supporting this idea, recordings from slices taken at various dorsoventral distances from the rhinal sulcus show that the intrinsic oscillations of MEC neurons scale with grid spacing (Figure 1.5d; Giocomo et al., 2007).

To understand the relationship between MEC grids and CA3 spatial maps, it is critical to examine them during remapping. Fyhn et al. (2007) made simultaneous recordings in MEC and CA3 while inducing varying degrees of hippocampal remapping. No changes in MEC were observed for rate remapping, but the grids underwent "realignment" contiguously with global remapping. We review entorhinal realignment further in Section 4.1.2, but it consists primarily of a randomized shift in the spatial phase of the grids (Figure 1.6a). The key characteristic of the relationship between realignment and global remapping is their contiguity: they occur simultaneously and on the same sub-minute timescale, and they are

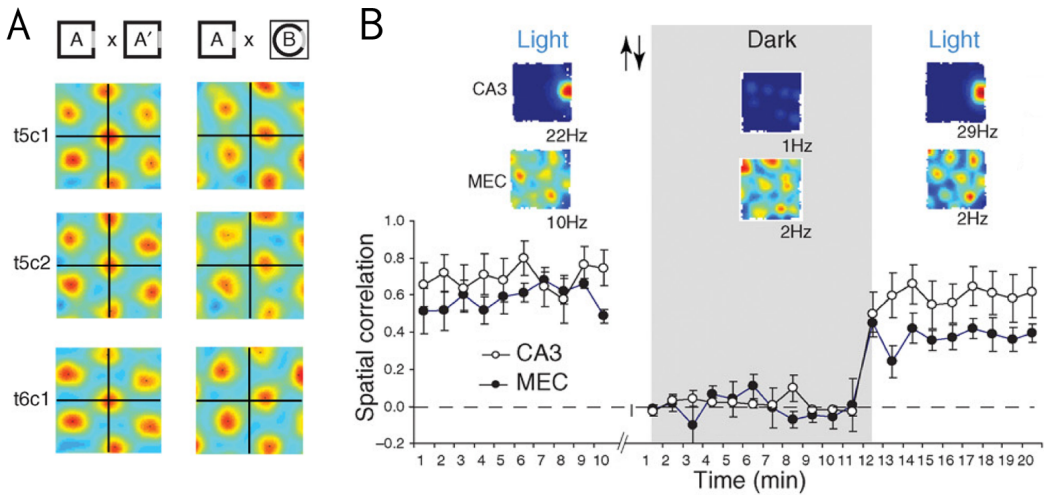

Figure 1.6 Simultaneous recordings in MEC and CA3 demonstrate cortical realignment contiguous with hippocampal remapping. A. Correlograms for example grid cells in one enclosure (left) show that they shift coherently in a different enclosure (right; same room). B. Rats form different spatial maps in the light and the dark. This test scenario shows a path-integration reset (double arrows) and turning off the lights decorrelates the light-based map in both MEC and CA3. Subsequently turning the lights back on restores the original maps on a sub-minute timescale. Adapted from Fyhn et al. (2007).

all-or-none processes. This contiguity is best illustrated by an experiment in which the rat switches spatial representations within the same environment depending on whether the lights are on or off (Figure 1.6b; Fyhn et al., 2007). Spatial correlations transition immediately once the switch is flipped. This is indicative of a direct mechanistic link between the two phenomena. Thus, we use our model of spatial map formation from Chapter 3 to explore this relationship between cortical realignment and hippocampal remapping (Chapter 4).

As discussed above, we can consider this stream of spatial information from MEC to dorsal CA3 as a putative rodent analog for the spatiotemporal context underlying human episodic memory. The area CA3 is unique in hippocampus for its dense associational synapses, which has inspired its extensive characterization as an autoassociative network (e.g., Marr, 1971; O'Reilly and McClelland, 1994). Au-

toassociation can perform retrieval through partial-cue pattern completion (Amit, 1989). However, since the recall of stored episodic memories requires that similar episodes must be encoded discretely, some source of pattern separation is required. DG has been recruited extensively in theoretical and computational models to perform this role by decorrelating entorhinal inputs (e.g., Teyler and DiScenna, 1986; Treves and Rolls, 1992; O'Reilly and McClelland, 1994; Norman and O'Reilly, 2003); this function is supported by recent experimental data (McHugh et al., 2007; Acsady and Kali, 2007). Further, Leutgeb et al. (2007) found that both DG and CA3, using simultaneous recordings, may contribute differentially to pattern separation: DG amplifies small amounts of environmental modification by modifying coincidence patterns; however, more substantial changes elicit the recruitment of a new and statistically independent cell assembly in CA3. Since the latter was not observed in DG, the CA3 mechanism may be due to changes in its direct entorhinal inputs. In other words, pattern separation in CA3 reflects global remapping and, therefore, the realignment of MEC grid-cell responses. Thus, realignment and global remapping can create orthogonal ensembles of active place cells that may enable the efficient encoding of episodes for recall.

1.4 Conclusion

We have briefly described two different mnemonic systems in the context of encoding/retrieval requirements and the division of labor within the MTL. The judgment of familiarity is a very high-capacity mnemonic capability in humans (Standing, 1973), but models of familiarity that integrate recollective processing greatly reduce that capacity. The neocortical inputs to familiarity networks are likely to be highly clustered, just as semantic networks show small-world structure. These

correlations improve feature recognition but interfere with recollection. This argues for the functional separation of the systems underlying recognition and recall. The unique structures of the hippocampal formation collect and associate parallel streams of objective and contextual information. The hippocampus seems to satisfy the computational constraints for efficiently forming the distinct orthogonal codes necessary for the efficient retrieval of episodic memory. The rodent system of spatial representation consisting of medial entorhinal cortex, based on randomly arranged grids, and its hippocampal targets provides an important model for understanding the recall of life experiences.

Chapter 2

Semantic Space and Frequency Effects in Recognition Memory

In the room where I work, I have a chalkboard, and as I'm going along, I write the made-up words on it. A few feet from that chalkboard is a copy of the full 20-volume Oxford English Dictionary, to which I refer frequently as a source of ideas and word roots. Whenever I get distracted or bored, my eyes wander over to that chalkboard and I read the words. Some of them grow on me, and others annoy me. I attack the latter with eraser and chalk, and keep nudging at them until I like the way they look and sound. Others never make the cut at all and simply get erased. Perhaps one day I will sell these on eBay to RPG players who need names for characters or alien races.

—Neal Stephenson, on creating new words

Summary

The word frequency effect (WFE) in recognition memory refers to the finding that rarer words are better recognized than more common words. We demonstrate that a familiarity discrimination model operating on data from a semantic word-association space yields a robust WFE in data on both hit rates and false-alarm rates. Our modeling results suggest that word frequency is encoded in the semantic structure of language, and that this encoding contributes to the WFE observed in item recognition experiments. ♦ The work presented in this chapter was published previously (Monaco et al., 2007).

2.1 Introduction

Old–new item recognition is the task of deciding whether or not test items were presented on a previous study list. Performance is quantified as the probability of old responses to (old) study items (hit rate, or HR) and to (new) nonstudy items (false-alarm rate, or FAR). One of the most prominent phenomena observed in this task is the word frequency effect (WFE): rare or low-frequency (LF) words are better recognized than common or high-frequency (HF) words (Schulman, 1967; Shepard, 1967). The recognition WFE is a mirror effect (Glanzer and Adams, 1985, 1990): it consists of an HR effect and an opposite but approximately equal FAR effect. The cause of the WFE and other mirror effects has been the subject of extensive study but no consensus view has been established (e.g., Murdock, 1998; Stretch and Wixted, 1998; Reder et al., 2000).

Both single- and dual-process models have been proposed to explain the WFE. The former perform familiarity discrimination (FD) based on similarity measures such as global feature matching. These models typically require some additional

transformation, such as log-likelihood computation, to achieve the required symmetry between old- and new-item familiarity distributions (Murdock, 1998). To explain the WFE, certain differences between LF and HF words must be assumed. These may include the modulation of attentionally marked features (Glanzer et al., 1993), diagnostic content (Shiffrin and Steyvers, 1997), or representative feature variability (McClelland and Chappell, 1998). In the end, such models produce a unidimensional scalar value for the strength, or familiarity, of a given stimulus that allows further analysis with signal-detection theory. HR and FAR calculations can be made by integrating thresholded familiarity distributions, and threshold-independent performance may be quantified with receiver operating characteristics (ROCs; see Wickens, 2002). Dual-process models, however, rely on differential contributions of recollective and familiarity-based processes to explain the performance differences. Recollection, a recall-like process, is characterized as less error-prone than a global-matching familiarity process (Guttentag and Carroll, 1997; Reder et al., 2000).

As discussed in Chapter 1, electrophysiological studies in monkeys have shown perirhinal cortex (PER) to have a central role in the stimulus-specific familiarity response in the MTL processing of recognition (Miller et al., 1991; Li et al., 1993; Xiang and Brown, 1998; Brown and Bashir, 2002). Theoretically, it is known that a familiarity signal can be read out from a simple autoassociative neural network by computing its internal energy (Amit, 1989). Indeed, the evaluation of network energy may approximate the familiarity signal evident in perirhinal neurons (Bogacz et al., 2001a; Brown and Bashir, 2002) and has been used to determine theoretical limits on recognition capacity (Bogacz et al., 2001b; Bogacz and Brown, 2002). Thus, we set out to create a recognition model that uses network energy as a readout of stimulus familiarity. For this purpose, we used input vectors from

a word association space (WAS; Steyvers et al., 2004). The WAS is an empirical model of semantic similarity based on normative data from free-association experiments (Nelson et al., 2004). Simulating old–new recognition experiments with this model, we found that word frequency produces discriminable signal distributions such that LF words tend to be more familiar than HF words. Further, coupling this output with a particular decision-making strategy exhibited a WFE mirror effect. These results have novel implications for the roles of distinct retrieval processes in recognition memory.

We present a simple item-recognition model where the familiarity of a probe stimulus is read out as the internal energy of a network trained on a set of activity vectors corresponding to WAS word representations. This is coupled with an experimental protocol emulating a typical word recognition experiment (see Methods: Experiment Simulation). Importantly, all study and test words are trained initially, and then followed by retraining of the study list. Retraining corresponds, here, to the subject having recently experienced a word in the context of an experimental study list.

2.2 Model

2.2.1 Hopfield network energy as a familiarity signal

In the Bogacz et al FD model, item vectors are associatively encoded into a Hopfield network (Hopfield, 1982). The familiarity signal is simply the internal energy of the network when activated with a probe stimulus (Bogacz et al., 2001a). Hopfield networks are fully-connected recurrent networks of binary units. The network weights are trained on an input set ξ_N^P of P N–dimensional activity vectors

such that $\xi_i^\mu \in \{-1, +1\}$ for all $i \in \{1..N\}$ and $\mu \in \{1..P\}$. That is, each unit is either active $(+1)$ or inactive (-1) for a given input vector. If we denote the weight matrix as $W = [w_{ij}]_{i,j=1}^N$, its elements are computed according to an associative Hebbian learning rule:

$$w_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu = \frac{1}{N}\, \xi_i \cdot \xi_j, \text{ for } i \neq j, \tag{2.1}$$

where $w_{ii} = 0$ for $i \in \{1..N\}$. Once trained, we are only interested in the internal energy of the network when presented with a given stimulus, so no network dynamics are involved here. This internal energy calculation is distinct from recollective processes that use some form of network relaxation to fully recall the features of stored items (Amit, 1989). For a probe stimulus vector $X = [x_i]_{i=1}^N$, the internal energy[1] is computed as

$$\mathbf{E}(X) = -\frac{1}{2} \sum_{i=1}^{N} x_i \sum_{j=1}^{N} x_j w_{ij} = -\frac{1}{2} X W X^T. \tag{2.2}$$

Note that more familiar stimuli will have *lower* energies than less familiar stimuli. A probe X will thus be associated with the familiarity quantity $\mathbf{E}(X)$ for a network trained on a given input set. In this form, the only free variables of the FD process are the size of the network, N, and the set of input vectors, ξ_N^P. FD such as this is more efficient and has a much higher capacity than associative recall. Allowing an error rate up to 0.01, the recall capacity of the network is $0.145N$ (Amit, 1989) whereas its recognition capacity is $0.023N^2$ (Bogacz et al., 2001b).

2.2.2 Inputs isomorphic to semantic similarity space

Recognition models operating on correlated input spaces (Bogacz and Brown, 2003; Norman et al., 2005) have been studied that benchmark behavioral data (Norman

[1]Here, we let $\mathbf{E}(\cdot)$ be the function mapping a stimulus or set of stimuli to an energy value or energy distribution, respectively. Statistical expectations are noted by $\langle\cdot\rangle$ brackets.

and O'Reilly, 2003). However, recent empirical models of semantic space, such as the WAS model of Steyvers et al. (2004), provide a basis for constructing an input set with a similarity structure derived from behavioral word-association data. Given that words are the stimuli most often used in human recognition-memory studies, it is instructive to utilize input vectors whose similarity relations approximate the lexical-semantic space of English speakers. The basis for the WAS is a free association dataset containing the probabilities with which subjects named a given word as the first associate of a cue word (Nelson et al., 2004). These data can be taken as a measure of direct associative strength among over 5,000 words. Indirect, or second-order, associative strengths can also be calculated from the dataset. To create the WAS, singular-value decomposition (SVD) was applied to these direct and indirect associations so as to place words in a reduced 400-dimensional space. This was constrained so that the cosine between any two word vectors[2] reflects their mutual associative strength. Strongly associated words have $\cos(\theta)$ values approaching 1, whereas semantically unrelated words have values approaching 0. The dimensional reduction revealed latent, higher-order semantic relations within the dataset. Importantly, 400 dimensions was found to be the lowest dimensionality that remains highly predictive of experimental data such as free-recall intrusion rates, extralist cued recall, and semantic similarity ratings in recognition (Steyvers et al., 2004). The resultant WAS shares gross structural characteristics, small-world but not scale-free, with other semantic networks (Steyvers and Tenenbaum, 2005). Thus, it is now possible to operate on input vectors whose similarity relations approximate the lexical-semantic space of English speakers.

We can now ask whether common and rare words differ in their similarity

[2]A vector cosine is the inner product between V and U normalized to $[0, 1]$, $\cos(\theta_{V,U}) = (V \cdot U)/(\|V\|\|U\|)$.

structure. To do this, we used a set of 1,748 WAS vectors for which we have the associated Kučera-Francis word frequency (WF; see Kučera and Francis, 1967). In the cosine similarity matrix for the 100 most common and the 100 rarest words in the set (Figure 2.1a), it is evident that common words tend to be similar to other common words and rare words tend to be similar to other rare words. Similarities between rare and common words tend to be lower than similarities within frequency groups. An intuitive reason for such differential encoding of frequency is that rare words tend to have a single definition while common words may have many definitions and usages. This reasoning predicts that rare synonyms will be clustered in semantic space, whereas more common synonyms will be placed at a semantic "centroid" of multiple distinct meanings. That is, LF words will be encoded into clusters and HF words will tend to occupy the space between such clusters. We will refer to this as the "tight clustering" hypothesis for LF words.

2.2.3 Interpreting the recognition model

The two functional components of our FD model are the Hebbian learning of a Hopfield network and the dimensional reduction of a word association matrix. These serve, respectively, as the FD mechanism and the semantic input space. In considering this combination, we have to interpret the necessary combination of the assumptions inherent in both. We must properly frame the limitations of the results and emphasize that they comprise, at most, a high-level explanation. A simple network computation on a carefully chosen input set will not explain the intricacies of human recognition memory for semantic stimuli, yet may provide insight into some aspect thereof. We argue for the specificity and functional plausibility of the model components as well as their composition.

First, we assume that employing the energy (Equation 2.2) of a trained Hopfield

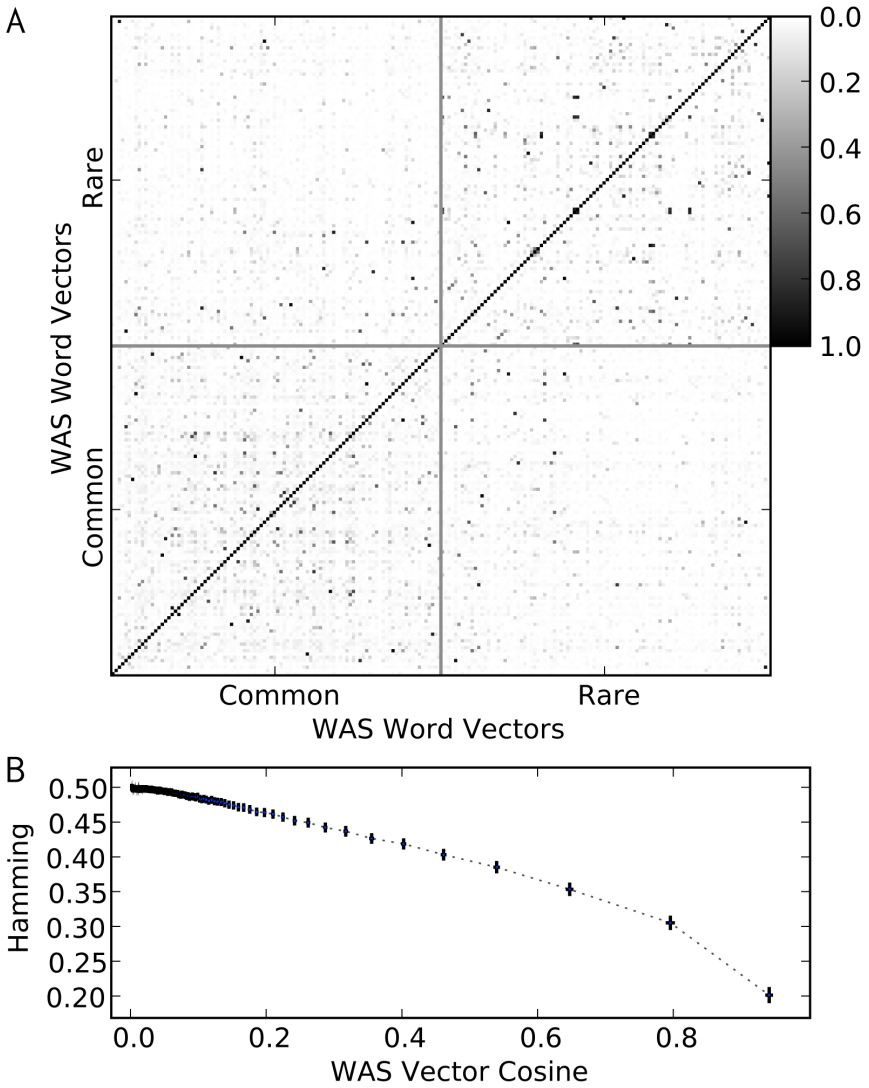

Figure 2.1 Similarity structure of the word-association space and binarization. A. Similarity matrix of the WAS vectors for the 100 highest and lowest frequency words in the set. The color of each pixel denotes the cosine of the angle between the *i*th and *j*th vectors. The lower-left and upper-right quadrants represent the cosine similarity among pairs whose members are both HF or LF words, respectively. The white diagonal signifies identity. The symmetric off-diagonal quadrants represent cosines between HF and LF words. B. Normalized Hamming distances between binarized WAS vectors decrease monotonically with the cosine of the corresponding WAS vectors. Every point represents the mean and error ($\alpha = 10^{-5}$) for each bin in a cosine-sorted partition (600 bins) of all vector pairs.

network (Equation 2.1) as a familiarity signal captures some salient characteristic of familiarity processing in the perirhinal cortex. Bogacz et al. (2001a,b) provide support for this assumption by arguing from a standpoint of functionality and efficiency as well as from modeling results. Also, neurons responding differentially to familiar stimuli have been found consistently within monkey PER (Miller et al., 1991; Li et al., 1993; Sobotka and Ringo, 1993). This difference is characterized by a reduction in stimulus-induced activity for familiar stimuli and rapid familiarity discrimination (on the order of 100 ms), but neural responses for recency and novelty have also been found (Fahy et al., 1993; Xiang and Brown, 1998; Brown and Bashir, 2002). Despite this functional diversity, only familiarity-sensitive neurons are considered here. Further, evidence from ablation and impairment studies indicate that the PER acts independently from other inferotemporal (IT) mnemonic systems such as that of the hippocampal formation (Gaffan, 1994; Murray and Bussey, 1999; Aggleton and Brown, 1999). This independence suggests that PER is the site of the neural substrate for familiarity judgments (for reviews, see Yonelinas, 2002; Rugg and Yonelinas, 2003).

Autoassociative neural networks such as the binary Hopfield network produce stimulus-dependent attractors (Hopfield, 1982). Reading out the internal network energy as a stimulus-familiarity signal is much more efficient than involving recollective processes (Bogacz et al., 2001a; Bogacz and Brown, 2003), which typically involve relaxing the network to reconstruct an attractor state (Murdock, 1982; Humphreys et al., 1989). For random vectors, this FD process has a very high storage capacity, as does human memory (Standing, 1973), and enables a rapid network response. It is also more robust than other network architectures; e.g., encoding via feedforward competitive synaptic processes can exhibit forgetting after a relatively small number of subsequent stimuli (Sohal and Hasselmo, 2000). Bo-

gacz et al. (2001b) use both a Hopfield network and a multilayer spike response model to argue that perirhinal FD neurons may form an autoassociative network in order to exhibit such efficiency and robustness. From this, we posit that it is reasonable that the energy computation of a Hopfield network is, at least, a useful abstraction of the FD processing performed by PER neurons.

Second, we assume that the WAS is at least approximately isomorphic to the space of neural representations of the semantic features of words for speakers of English. This amounts to the assumption that behavioral associativity reflects the neural encoding of semantic similarity. The WAS consists of transformed statistical behavioral data from 6,000 subjects, and as such can only be inferred to resemble the structure of semantic space for a given subject. That the WAS serves well as a predictive model for known human memory effects recommends it as a useful inference of semantic space. Further, the WAS has structural characteristics consistent with being isomorphic to real semantic representations. Thus, this assumption is assuredly a simplification, but it is likely to yield salient semantic information. Even though the WAS vectors were binarized in order to be used as proper inputs to the Hopfield network, Figure 2.1b shows that this transformation preserves the gross structure of the space.

Third, we assume that the semantic WAS vectors serve as appropriate inputs to the FD model of PER. This assumption allows us to posit the combination of the two components as a unified model of recognition memory. Supporting this, several clinical studies indicate a role for PER in associative memory for semantic content and lexical processing (for review, see Murray and Bussey, 1999). Further, neurons in the perirhinal and other IT areas in monkeys demonstrate the ability to represent abstract object categories (Erickson et al., 2000; Miller, 2000; Miller et al., 2003). Such abstraction is a hallmark of semantic information processing and

indicates that the PER has access to semantic features among its inputs. Thus, the semantic features of words that are presented to subjects as stimuli in recognition experiments are presumed to be accessible from PER.

Finally, although we investigate the recognition WFE with this model, it can only describe effects due to familiarity processing of semantically structured input data. There are certainly non-semantic contributions to the WFE that are not within this scope; e.g., context variability, orthographic and phonological features (Steyvers and Malmberg, 2003; Malmberg et al., 2002). Qualified as such, we will refer to this bipartite recognition model as WAS–FE.

2.2.4 Hopfield energy and semantic attractors

If a meaningful stimulus is ultimately represented as a binary pattern of activation across N perirhinal neurons, then we can think of this stimulus as an N-dimensional vector of features that are either present in the stimulus ($+1$) or not (-1). In Hopfield learning (Equation 2.1), these component features are pairwise associated according to their correlation: the strength of the synapse between two neuronal units is directly and linearly related to the number of patterns for which the units carry the same activity. Synaptic weights are simply inner products of across-pattern activity vectors. Internal network energy (Equation 2.2), then, is an outer product measure of how well the pairwise bit structure of a given activity vector aligns with the pairwise correlations stored in the weights of the network. This is opposed to recognition models based on summed similarity or global matching (Shiffrin and Steyvers, 1997; Zaki and Nosofsky, 2001; Kahana and Sekuler, 2002; Kahana et al., 2005). However, a summed-similarity recognition model using inputs derived from perceptual preprocessing of natural stimuli has been able to match experimental similarity and recognition data (Lacroix et al.,

2006).

Small-world structure is characterized by short minimum-path lengths but also by hub-like connectivity (Watts and Strogatz, 1998). Considering the WAS as a small-world network (Steyvers and Tenenbaum, 2005), there must be subsets of vectors that significantly share pairwise activity. Each of these groups, or clusters, will bias those synaptic weights corresponding to their respective set of shared features. Clusters of feature-sharing vectors will form attractors commensurate with their size and mutual similarity. Thus, a probe vector may yield a low energy by matching features characteristic of different attractors in the network: there is a combinatoric aspect to the diversity of such "spurious" attractors (Amit, 1989). The strongest attractors, though, will correspond to groups of words with substantial semantic similarities.

2.3 Methods

2.3.1 Signal detection

The noise in the synaptic weights of the network are due to mutual interference between stored patterns, and depends on both on the number of patterns (P) and the variability of their correlational structure. This synaptic noise is translated into randomness in the internal energy computation (Equation 2.2) and, therefore, the familiarity measurement of a given probe. For WAS–FE to serve in a recognition experiment, a binary old–new decision must be made from this noisy scalar output. These conditions satisfy the assumptions necessary to assess recognition performance using signal detection methods (Wickens, 2002).

A decision threshold, or criterion, can be used to efficiently decide if a probe

stimulus is familiar or not. For a criterion λ, a vector X is determined to be old if $\mathbf{E}(X) < \lambda$, otherwise it is judged new. The distribution of energies from trained vectors is distinct from that of untrained probes. Consider a random and unbiased set ξ_N^P of stored vectors. The distribution of synaptic weights in W can be approximated by a Gaussian distribution with $\mu_W = 0$ and $\sigma_W^2 = P/N^2$. The energy distributions for both untrained probes and stored vectors have $\sigma_{\mathrm{E}}^2 = P/2$. The expected energy value of an untrained probe X is $\langle \mathbf{E}(X) \rangle = 0$ while that of a stored vector ξ^μ is $\langle \mathbf{E}(\xi^\mu) \rangle = -N/2$. Here the logical decision criterion would be $\lambda = -N/4$, the midpoint between the old and new energy distributions. This is the criterion employed by Bogacz et al. in their signal–noise analysis of capacity 2001b; 2002. For the semantically structured inputs considered here, λ is chosen as the midpoint between the empirical means of the distributions. To determine HRs and FARs, we used a WF-based multiple-criterion decision strategy (see Discussion) with means-based thresholds,

$$\lambda_i = \left[\langle \mathbf{E}(\Phi) \rangle + \langle \mathbf{E}(\Lambda_i) \rangle \right] / 2, \tag{2.3}$$

where $\mathbf{E}(\Phi)$ is the energy distribution of the reference pool. The performance of the model is assessed by calculating HRs and FARs. The HR is the fraction of stored vectors with $\mathbf{E} < \lambda$, while the FAR is the probability for an untrained vector to have $\mathbf{E} < \lambda$. The discriminability between the old ($\mathbf{E}(\xi^{\mathrm{o}})$) and new ($\mathbf{E}(\xi^{\mathrm{n}})$) energy distributions is computed as the distance between means in standard deviations,

$$d'(\mathbf{E}(\xi^{\mathrm{n}}), \mathbf{E}(\xi^{\mathrm{o}})) = \frac{\langle \mathbf{E}(\xi^{\mathrm{n}}) \rangle - \langle \mathbf{E}(\xi^{\mathrm{o}}) \rangle}{\sqrt{(\sigma_{\mathrm{n}}^2 + \sigma_{\mathrm{o}}^2)/2}}. \tag{2.4}$$

For the experimental data in Figure 2.4b, d' was calculated based on an unbiased estimator assuming underlying Gaussian distributions: $\hat{d'} = z(\langle HR \rangle) - z(\langle FAR \rangle)$. ROCs are constructed by plotting HR against FAR for a range of possible decision criteria. Thus, they provide a criterion-independent assessment of how well the

familiarity signal is discriminated—one that is especially informative in the case of non-Gaussian distributions. Better performance is indicated by an ROC curve farther from the chance function (where HR = FAR and $d' = 0$) in the direction of higher HRs and lower FARs.

2.3.2 Simulation of the recognition experiments

In an old–new item-recognition experiment, the subject studies a list of known items from a training set. At test, the subject is shown a list of probe items, some of which had appeared in the training set (old items) and others which had not (new items). The task is to judge whether each item is old or new.

Experimental subjects here are defined by Θ, the random subset of word vectors on which the Hopfield network is initially trained (Equation 2.1). Each vector in Θ is associated with a WF value corresponding to the word that it represents. Using these frequencies to index the vectors, Θ is sorted and evenly partitioned into 6 bins with Θ_1 containing the highest-frequency subset and Θ_6 containing the lowest-frequency subset of words. The other four subsets contained word vectors of intermediate WF values. The Θ_i are equally sized to within one vector due to rounding. The study list for the task is defined by Λ, which is a random subset of Θ such that an approximately equal number of study items are chosen from each WF bin. That is, Λ comprises random subsets $\Lambda_i \subset \Theta_i$, for $i \in \{1..6\}$, such that the length of the study list is $L = \sum_{i=1}^{6} |\Lambda_i|$ where $|\Lambda_i|$ is the number of vectors in Λ_i. These WF study bins were equally-sized to within one vector due to rounding. The study list is presented to the model by retraining the network on all the study vectors. If ξ_Λ is a matrix containing the study vectors row-wise, then W is updated

by computing

$$W \to W + \frac{1}{N}\,\xi_{\Lambda}^{T}\xi_{\Lambda} \tag{2.5}$$

and then zeroing out diagonal terms. This procedure is analogous to strengthening the pre-existing neural representation of the items in a study list attended to by a subject. Specifically, this operation doubles every weight component resulting from the initial training of Λ as part of Θ. All items are studied equally. The training-set vectors not chosen for the study list composed the reference pool, Φ. Therefore, the size of the pool is the size of the training set Θ minus the study-list length $|\Lambda|$. The items in the study list and the reference pool serve as the old and new probes during test, respectively. We calculate the internal energy (Equation 2.2) of each vector in Θ using the updated weight matrix. We then compare the resulting sample energy distributions by calculating d' distances (Equation 2.4), ROC curves, HRs and FARs. This process, starting with a new random Θ chosen from our binarized WAS, is repeated for 2,000 trials. That is, one trial here is analogous to a new subject performing a single recognition task. Across-trial means and confidence intervals were computed for Figure 2.3 and Figure 2.4. Representative energy histograms were created by accumulating energy vectors across trials, computing counts for 100 equally spaced bins across the range of energies, and the scaling the bin frequencies.

2.3.3 Neighborhood measures in association spaces

We performed two basic neighborhood analyses. First, we computed the WF-dependence of the mean number of neighbors in WAS space. Consider the metric $d_{\cos} = 1 - \cos(\theta)$ so that close neighbors have a cosine approaching 1 and a $d_{\cos}$ approaching 0. For every WAS vector, our algorithm traversed the range of possible

$d_{\cos}$ from 0 to 1 while counting the number of vectors within that distance from the given vector. This cumulative measure constitutes a population count for neighborhoods of increasing radii in WAS space. The mean populations were computed for every $d_{\cos}$ radius for each of six WF classes (Figure 2.5a). Second, to address the question of the WF composition of neighbors, we employed a $\cos(\theta)$-weighted frequency measure. We calculated the quantity,

$$\nu(\xi^{\mu}) = \sum_{\varphi=1}^{1,748} \left(\frac{\xi^{\mu} \cdot \xi^{\varphi}}{\|\xi^{\mu}\| \|\xi^{\varphi}\|} \right) f_{\mathrm{KF}}(\xi^{\varphi}) - f_{\mathrm{KF}}(\xi^{\mu}) \tag{2.6}$$

for all WAS vectors ξ^{μ}, where $f_{\mathrm{KF}}(\cdot)$ is the function mapping a vector to its associated Kučera-Francis WF value. This measure quantifies the expectation of the WF for neighbors of a given vector. The distributions of these convolutions for the word vectors of each of the six frequency classes are shown as 15-bin histograms in Figure 2.5b.

2.4 Results

Initially, we binarized[3] the WAS vectors in our wordset. This allows proper operation of the Hopfield learning rule (Equation 2.1) and energy computation (Equation 2.2). The normalized Hamming distances between these binary vectors decreased monotonically with cosine similarities of the corresponding continuous WAS vectors (Figure 2.1b). This indicates that the binarization significantly preserves similarity relations between vectors.

[3]The elements of WAS vectors are symmetrically distributed around zero, so that taking the sign of each vector element produces a set of binary vectors with, on average, unbiased activity levels.

2.4.1 Random input case

In the initial item-recognition experiment we used an input set of unbiased, random vectors. The energy distributions for old and new items were binomial with means of -399 and -199 (Figure 2.2a, left column), matching theoretical means[4] of $\mu_\Lambda = -400$ and $\mu_\Phi = -200$. Study lists consisted of 100 vectors ($L = 100$) and results using two reference pool sizes are shown in Figure 2.2a: $P = 400$ (top row) and $P = 1600$ (bottom row). We refer to these as the low-load and high-load training conditions, respectively. They demonstrate the effects of adding noise to the network in the form of additional stored vectors. In all cases, the old and new distributions have equal variance. The low- and high-load conditions had standard deviations s_{E} of 20 and 31, respectively. This increase in spread decreases d' from 10.1 to 6.5 (Equation 2.4) in the high-load condition, however both values indicate perfect discrimination. Although we could have degraded the model's performance by reducing the magnitude of the weight update for study list items (Equation 2.5), these results serve as an input control for the simulations below.

2.4.2 Semantic inputs from word association spaces

The energy distributions resulting from the semantic input set (Figure 2.2a, right) differ substantially from the random input condition. The distributions are non-normal, negatively skewed (i.e., biased toward increasing familiarity), and their statistics have changed significantly. The means are lower than those in the random input case. In the low-load condition, mean old and new energies are -524 and -326, respectively, and the high-load case shows -777 and -578, respectively.

[4]These means are offset $-N/2$ due to the initial training of both study and test vectors (see Methods: Experiment Simulation).

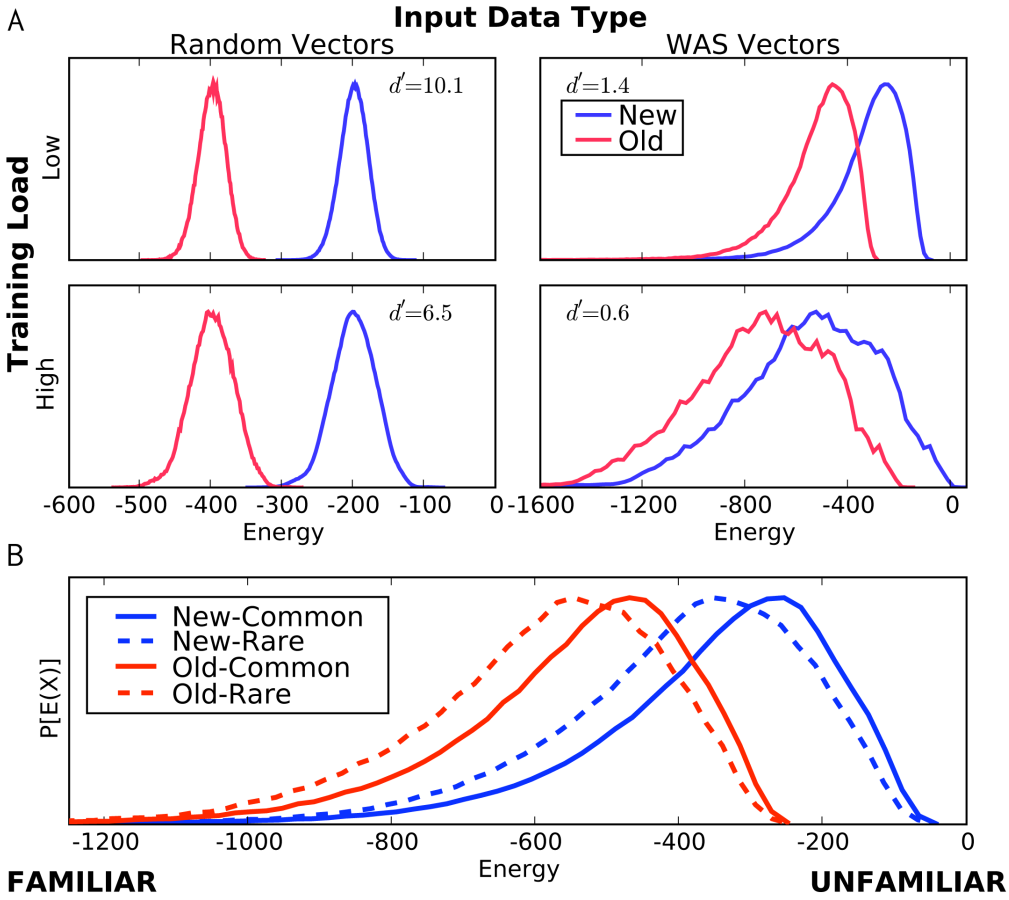

Figure 2.2 Effects of semantic input vectors and training load on resultant energy distributions. A. Increasing training load for both random (left) and semantic (right) vectors increases overlap between energy distributions (100 study items, 400/1600 new items for low/high (top/bottom) training load). The energies for semantic inputs, however, have load-dependent means, non-Gaussian distributions, and worse discriminability than in the random case. B. WF-sorted partitioning of word vectors results in discriminable familiarity distributions (150 study items, 600 new items). The LF distributions are more familiar than HF words for both old and new items. The "rare" and "common" bins here are the least and most frequent thirds of the lists, respectively.

So, not only are the distributions exhibiting enhanced familiarity, the means are load-dependent. The more semantic vectors we store in the network, the more negative the energy distributions become. Further, they exhibit much lower d' distances than in the random case. The discriminability as measured by d' decreases

from 1.4 at $P = 400$ (Figure 2.2a, top right) to 0.61 at $P = 1600$ (Figure 2.2a, bottom right). This 57% reduction in separation compares to a corresponding 36% decrease for random inputs. Finally, the high-load condition produces energy distributions with noise-like irregularities that are not evident in the other cases. These were not investigated, but they may be the result of capacity effects or structural heterogeneity of the input space.

Statistical changes such as these could be expected for any sufficiently non-random input set. However, there are systematic differences in the energy distributions among WF classes. We found that vectors representing LF words tend to have lower energies, and thus enhanced familiarity, than those of HF words (Figure 2.2b). This was observed for frequency classes in both old and new energy distributions. Figure 2.2b shows the distributions for the thirds of the study list and reference pool with the highest and lowest frequencies. This effect of increasing familiarity with decreasing WF was observed robustly across the full range of possible list and reference pool sizes. For the data shown here, based on $L = 150$ and $P = 600$, all four distributions exhibited standard deviation $s_{\mathrm{E}} = 192$, and both WF-dependent effects were discriminable at $d' = 0.23$. This effect is present in the new distribution and, as such, does not depend on item study (Equation 2.5).

ROCs computed from the semantic familiarity distributions in Figure 2.2a are presented in Figure 2.3. The WF-dependence of the ROCs is shown for both the low-load (Figure 2.3a) and the high-load (Figure 2.3b) conditions. That is, for each training condition, the "common" and "rare" ROCs compare Λ_1 (old–HF) and Λ_6 (old–LF), respectively, to the reference pool. These are the distribution comparisons used to assess item-recognition performance (see below and Discussion). In both conditions, LF words yield better old–new discrimination than HF words. There are two load effects. First, the low-load ROCs (Figure 2.3a) indicate better

overall performance, evident as higher HRs and lower FARs, than the high-load ROCs (Figure 2.3b). Second, the WF-dependence of effect is greater in the high-load than in the low-load condition. That is, the ROCs in Figure 2.3b are more separated than those in Figure 2.3a. Both load effects are a result of the increase in energy variance and decrease in d' distances evident in Figure 2.2a. In the low-load condition, the d' distances are 1.2 and 1.5 for common and rare words, respectively. For high-load, the relatively low d' of 0.34 for common words more than doubles to 0.78 for rare words. This Hopfield FD model has a theoretical recognition capacity of 3.7×10^3 random vectors (Bogacz et al., 2001b). Here, storing 1.7×10^3 semantic vectors is severely detrimental to FD performance, indicating that the correlations inherent in the semantic inputs reduce the effective capacity of the network (Appendix C of Bogacz and Brown, 2003).

2.4.3 Word-frequency effects in recognition

Calculating HRs and FARs requires a decision criterion. Here, we employ a multiple-criterion decision paradigm that makes comparisons between the entire reference pool and each individual study-list frequency class. For each comparison, the decision criterion λ_i is chosen as the midpoint between the empirical means of the Λ_i and reference pool distributions (Equation 2.3). We observed the WFE mirror effect across the possible range of the list lengths for the study list and reference pool. We based recognition performance on a WF-based decision criterion (see Discussion). Mean HR–FAR trends for the low-load condition are shown in Figure 2.4a with d' distances. In this condition, HR decreases from 0.77 for the lowest frequency words to 0.73 for the highest frequency words. Similarly, the FAR increases from 0.18 to 0.22. Human recognition data collected by Schwartz et al. (2005) are shown for comparison in Figure 2.4b. The experimental d' distances

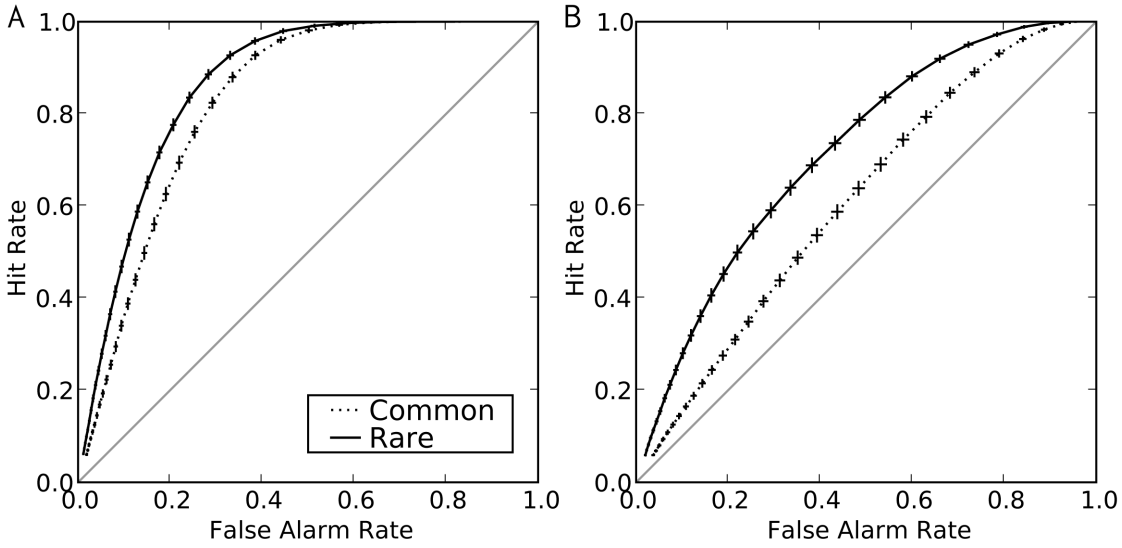

Figure 2.3 ROC curves for recognition. Operating characteristics for item-recognition performance under low (A; $P = 400$) and high (B; $P = 1600$) training load. The study list is composed of 100 items in both conditions. Performance across word frequency is assessed using a 6-partition of the study list indexed and sorted by Kučera-Francis frequencies. The common and rare ROC curves represent the performance of the highest and lowest frequency bins, respectively. The $P = 400$ case (A) demonstrates better baseline performance but a smaller frequency effect than the $P = 1600$ case (B).

(Figure 2.4b, bottom) were calculated using an unbiased estimator from detection theory (Wickens, 2002). The experimental HR decreases from 0.90 to 0.86, while the FAR increases from 0.077 to 0.13. This data approximately matches the trends observed in the model data.

Note, however, the differences in absolute magnitude of the HRs, FARs, and d' distances between the model and experimental data in Figure 2.4. The absolute d' distances could be manually tuned with the addition of a coefficient in the study rule (Equation 2.5), but we chose not to do this. Scaling up the model d' data would increase HRs and decrease FARs to better match the experimental data. For our purposes, it is sufficient that we observe a qualitatively correct WFE.

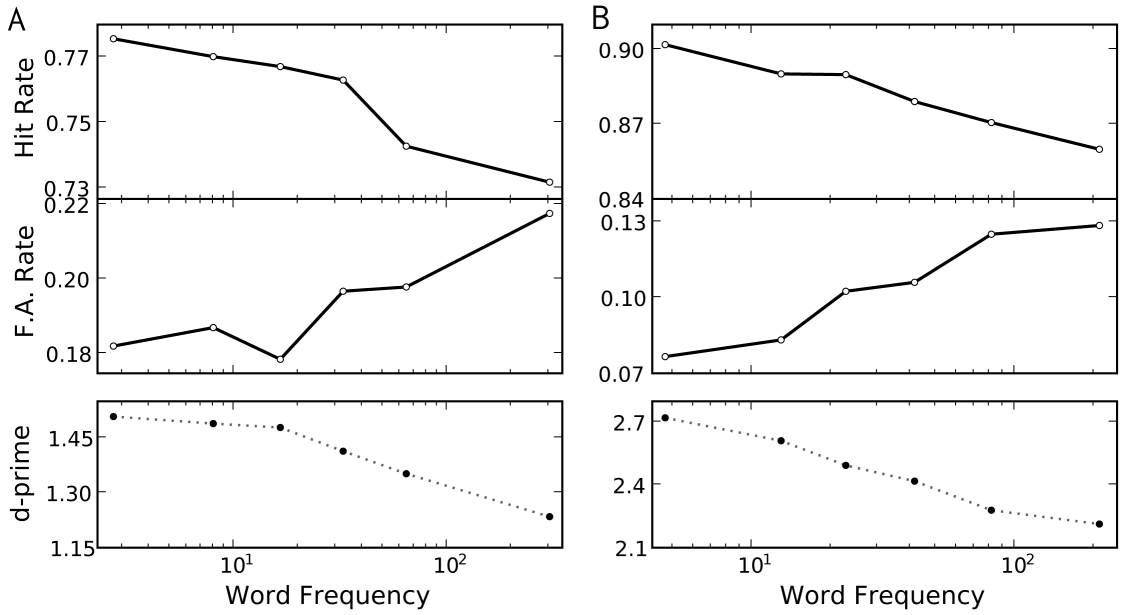

Figure 2.4 Mirror effect for word-frequency groups. Word frequency mirror effect from 2,000 trials of item-recognition experiment simulations (A; $L = 100$, $P = 400$) and as seen in human memory experiments (B; data from Schwartz et al. (2005)). The HR and FAR effects compose the mirror effect (top) and are due to changes in discriminability (bottom). Model HR, FAR, and d' data have 95% confidence intervals of mean $\pm$ 4.2×10^{-3}, 1.5×10^{-3}, and 1.2×10^{-2}, respectively. The discriminability of the experimental data was estimated from signal detection theory.

2.4.4 Differential clustering in semantic space

The above supports the tight clustering hypothesis for LF words, so we performed two simple neighborhood analyses of the continuous WAS dataset. Consider an even partition of the entire WAS such that each bin contains a distinct WF class. Figure 2.5a shows, for each of the WF bins, the mean neighborhood population counts for word-centered hyperspheres of varying radii in cosine space. We count all neighbors, regardless of word frequency. Counts are shown for radii up to $d_{\cos} = 0.062$, as that is sufficient to illustrate the WF-dependence of the number of close neighbors as $d_{\cos}$ approaches zero. The rarest words (blue solid line) have more neighbors on average than common words (red lines) for most of this range,

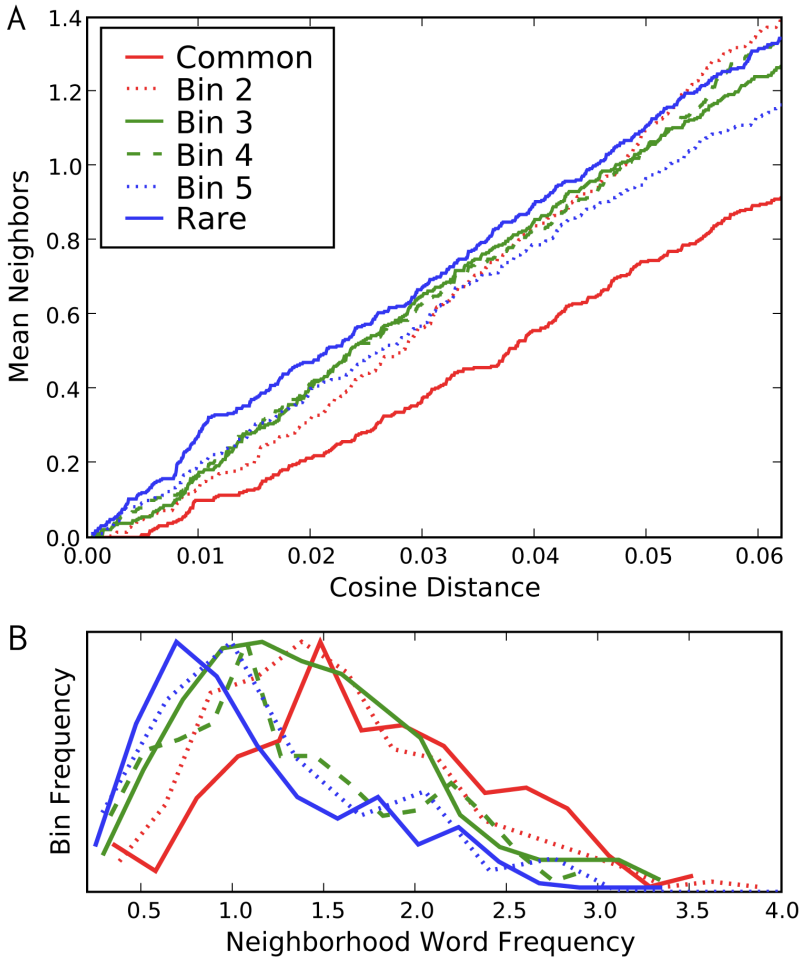

Figure 2.5 Word frequency-dependent clustering of word vectors in the WAS. Using a 6-partition of the word set, the mean population size for a neighborhood of a given radius in cosine space shows that rare words have more close neighbors than common words (A; the abscissa is $d_{\cos} = 1 - \cos(\theta)$). The WF composition of those neighbors is indicated by the relative distributions of WF–$\cos(\theta)$ convolutions (Equation 2.6) for different WF bins (B). The close neighbors of rare words tend to be other rare words.

and especially around $d_{\cos} = 0.01$ to 0.02. This indicates that LF words tend to have more close neighbors than HF words.

Next, Figure 2.5b addresses the WF-composition of those neighbors. For each vector ξ, we compute the quantity $\nu(\xi)$ as an average of the WF of its neighbors (Equation 2.6). The distributions of these values for each frequency class

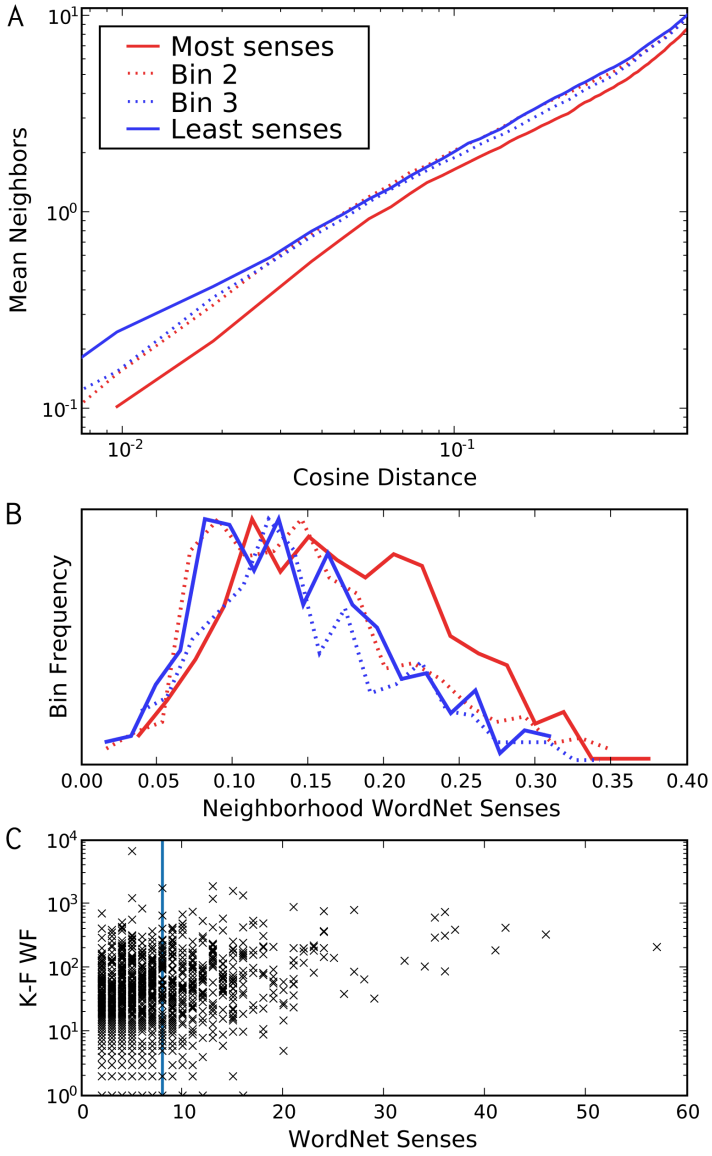

Figure 2.6 Clustering analysis redone in terms of the number of WordNet senses for a given word. A,B Instead of the K-F normative word frequency, the wordset used is the intersection of WAS words with K-F values and those with WordNet results. C. WordNet senses are plotted against WF for comparison, with the blue vertical line indicating the lowest number of WordNet senses for Bin 1, which contains the words with the highest number of senses.

are shown as 15-bin histograms in Figure 2.5b. The LF and HF distributions have means of 1.13 and 1.74, respectively, and a distance of $d' = 0.93$ standard devia-

tions. From Figure 2.5a, LF words tend to have closer neighbors than HF words. Given the null hypothesis that semantic similarity and WF are not correlated, the $\cos(\theta)$ coefficients in (Equation 2.6) would dictate that these distributions be orientated oppositely from those in Figure 2.5b. That is, even though LF words have more high-similarity neighbors – so that neighbor WF values are more strongly weighted – $\nu(\xi)$ is distributed to lower WF than those of common words. Thus, the dominant factor in these $\nu(\xi)$ averages must be the WF component, indicating that LF words are tightly clustered with other LF words, whereas HF words are more diffusely distributed.

2.5 Discussion

Here, we bring together a simple model of familiarity-based recognition (Bogacz et al., 2001a,b) and a recent model of semantic similarity (Steyvers et al., 2004) and demonstrate a word frequency effect. The only free parameters for the resulting model (WAS–FE) are the lengths of the study and test lists, across which the WFE is robust, only differing in magnitude. Notably, and as discussed below, the observed WFE is a mirror effect when decisions are determined using a stimulus-dependent criterion.

2.5.1 Effects of the semantic structure of cortical inputs

For random and unbiased pattern sets and probe vectors, the statistics of the weights and energies can be determined *a priori* (see Methods). Sample energy distributions resulting from the item-recognition experiment are shown in Figure 2.2a. The WAS-based semantic vectors are not activity biased, and the lists in each trial are unbiased with respect to WF. Training and testing with these mixed-list seman-

tic inputs produce energy distributions which are negatively skewed and exhibit statistics different from the random case (Figure 2.2a). The input-type effect evident in Figure 2.2a is attributable to non-random structure of the semantic input space. From the low means for the new distributions, we can infer that the vectors in the semantic input space tend to be near network attractors. The presence of spurious attractors, as well as learned attractors, further contributes to lower energies across the space. In fact, the large systematic decrease in probe energies indicates that much of the input space is likely spanned by basins of attraction. This is expected in a small-world network, because most vectors are part of the local clusters which give rise to the network attractors in the first place. The observed shape of the energy distributions, then, is a function of the number, density and spatial distribution of vector clusters in WAS.

Vectors populating a space with large-scale structure carry redundant information in their correlations. For the Hopfield FD mechanism, it can be shown that input redundancy reduces the effective capacity of the network (Appendix C of Bogacz and Brown, 2003). This is evident in the large drop in discriminability between the random and semantic input spaces and between the low- and high-load training conditions (see d' in Figure 2.2a). So, processing raw semantic information is inefficient, but results in much more realistic (i.e., measurably worse than perfect) recognition performance. Stimulus de-correlation is thought to occur downstream of IT cortex in the dentate gyrus (O'Reilly and McClelland, 1994; Kesner et al., 2000), so presumably the FD neurons in PER have access to the original stimulus features.

2.5.2 Word-frequency effects from Hopfield energies

The WFE of recognition memory is one of the most robust and extensively studied human memory effects (e.g., Schulman, 1967; Shepard, 1967; Glanzer and Adams, 1985; Guttentag and Carroll, 1994; Karlsen and Snodgrass, 2004). On the hypothesis that the WFE is supported by semantic coding differences and thus might be evident within WAS–FE, we sorted and partitioned the task lists into WF bins using a normative frequency measure (Kučera and Francis, 1967). These WF-differentiated bins resulted in separable energy distributions for both old and new lists (Figure 2.2b), an effect observed robustly across list length. Notably, this frequency effect is in the observationally correct direction of increased familiarity for rarer words. That is, as in Figure 2.2b, the energies for each "Rare" (LF) bin are distributed more negatively than those of the corresponding "Common" (HF) bin. This effect is robust across all free parameters in the model, which are just the study and test list lengths.

There must be some structural or statistical characteristic of the WAS underlying this effect: aside from the network mechanism, there simply is nothing else in WAS–FE to cause it. Specifically, the random input condition (Figure 2.2a) serves as control for the structure in the semantic inputs. Everything aside from semantic structure as represented by the WAS is controlled for in the random input case (Figure 2.2a). We performed two simple analyses of the original WAS vectors to support the hypothesis that LF words tend to be tightly clustered with other LF words. Cumulative population counts across cosine space (Figure 2.5a) show that LF words have more close neighbors than HF words. The structural hypothesis only makes sense if those neighbors tend to be other LF words and that the sparser neighborhoods of HF words tend to consist of HF words. This is because of the associative learning mechanism: when more pairwise correlations (i.e., semantic

features) are shared among a subset of input vectors, the corresponding network weights are strengthened. These stronger weights result in a stronger attractor and lower energies for vectors in the self-similar subset. Further, distributions of a neighborhood frequency average (Equation 2.6), demonstrate the required WF-dependence and separation to show that words tend to be co-located with neighboring words of the self-same WF class (Figure 2.5b). This WF-based coding scheme is evident despite the tighter LF clustering shown in Figure 2.5a: the closest neighbors of LF words contribute larger $\cos(\theta)$ factors to the summation (Equation 2.6) than HF words, so it follows that if clusters were heterogenous with respect to WF then the opposite tendency would be observed. These neighborhood effects support the tight-clustering hypothesis, that LF words tend to cluster with other LF words while HF words are coded more diffusely. Considering attractor formation on small-world inputs, this is sufficient cause for the type of relative familiarity differentiation observed in Figure 2.2b.

Among single-process recognition models, there has not been a consistent approach for the structural representation of semantic stimuli. For instance, the retrieving effectively from memory (Shiffrin and Steyvers, 1997) model assigns higher diagnostic content to LF words by spreading out the distribution of feature values for LF words. This is based on the assumption that HF words share features to a higher degree than LF words. However, the subjective-likelihood model (McClelland and Chappell, 1998) approaches WF differentiation by injecting more noise into the feature vectors of HF words to represent the higher degree of contextual variability for more frequent words. Notably, a normative measure of context variability has been shown to have a recognition mirror effect independent of WF (Steyvers and Malmberg, 2003). Lastly, the attention-likelihood theory (Glanzer et al., 1993; Malmberg and Nelson, 2003) does not rely on structural differences to

demonstrate the WF mirror effect. Instead, it employs the hypothesis that fewer features of HF words are attended to by the subject. This effectively reduces the semantic information in HF words, analogous to adding noise or placing vectors in smaller clusters. Algorithmically, these models combine feature matching with the computation (or estimation) of log-likelihood ratios.

One intuitive line of thought is that a semantic attractor represents an atom of semantic content, a "sense". LF words tend to be associated with a very small number of different senses. HF words, however, may be associated with many distinct word senses. Semantic encoding would then place HF words in the space between their several senses. This leaves most rare words proximal to strong semantic attractors, with corresponding high familiarity judgments, and more common words are placed away from these energy minima. To assess the validity of this idea, we repeated the neighborhood analyses for a WAS 4-partition based on the number of senses a word has in the WordNet database (Fellbaum, 1998). The cumulative population count Figure 2.6a shows a significant decline in close neighbors only for the most-senses words. The $\nu'(\xi)$ (i.e., neighborhood senses average) distributions Figure 2.6b show only the most-senses words have a slight tendency to be encoded with other high-senses words. Finally, plotting WF against WordNet senses Figure 2.6c reveals that only the most-senses words have any correlation with word frequency. This indicates that the number of word senses does contribute explanatory power to our structural observations, but this is limited to the HF/most-senses domain.

2.5.3 Decision process and performance quantification

The WFE is fundamentally a behavioral effect of recognition performance, so a decision-making process is needed. The human WFE is a mirror effect (Glanzer

and Adams, 1990; Glanzer et al., 1993), meaning that, for LF probes, subjects are better at both accepting targets as old and rejecting lures as new. Accordingly, our decision process must be able to demonstrate both the HR and FAR aspects of the LF enhancement if WAS–FE is to have any explanatory power. Many different decision processes could be devised, but we will explore what would be necessary with the restriction of a simple process within the scope of and commensurate with the results presented so far.

We classify WAS–FE as a single-process signal detection model of recognition: LF and HF stimuli are not processed differently and a single scalar energy value is the only output. This familiarity signal is noisy and a decision must be made whether a given probe was studied (old) or not (new). We consider a simple threshold process (see Methods: Signal Detection) with a decision criterion below which a probe is judged old, otherwise new. For simplicity, we will consider, as decision criterion, the midpoint of the empirical means of the energy distributions. Similarly, but for random vectors, Bogacz et al. (2001a,b) employed the midpoint of the theoretical means. The question becomes that of which two distributions, in particular, are being compared in the decision process. The answer to this affects the interpretation of WAS–FE as a candidate explanation for the recognition WFE.

The signal detection comparison here is between a study list (Λ) and a reference pool (Φ), either of which may be broken down in WF classes. Thus, there are 2×2 possible comparisons: the WF bins or aggregate study list against the WF bins or aggregate reference pool. In the notation used above and in Methods: Experiment Simulation, these are Λ_i–Φ_i, Λ_i–Φ, Λ–Φ_i, and Λ–Φ, respectively, where i traverses WF bins. The Λ–Φ comparison is not WF-dependent and thus meaningless in terms of the WFE. With a means-based criterion, the Λ_i–Φ_i comparison will not produce either component of the mirror effect because the WF-dependence of both distri-

butions is the same. A fixed, WF-independent criterion could be used, but the resultant FAR effect would not "mirror" the HR trend. This is typical of the fundamental difficulty with single-process signal detection models, as the familiarity effect needs to be reversed for new items to achieve a mirror effect (Glanzer et al., 1993). For instance, the attention-likelihood theory of the WFE mirror effect uses a log-likelihood ratio to bring about this required symmetry (Murdock, 1998).

We are left to consider decisions involving a mixed comparison: WF bins of one distribution against the aggregate of the other. The Λ–Φ_i comparison falters on two counts. First, if only one distribution is going to receive the benefit of WF information, it does not make sense for it to be the distribution for items that have not been recently experienced. Second, it results in a performance decrease with rarity because the increasingly negative distributions have more overlap with the Λ distribution. The Λ_i–Φ comparison addresses both counts: cognitively and intuitively, it makes sense that the subject has information regarding the WF classes of recently studied stimuli; and performance increases with rarity because the old distributions are farther from the Φ distribution. There are different possible forms for a stimulus-dependent criterion shift, but most allow that the criterion must increase in the signal direction "with the memorability of old items" (Hirshman, 1995). Thus, we can tentatively delimit certain requirements for both the discimination comparision (Λ_i–Φ) and the decision criterion (Equation 2.3). The resultant WFE has mirrored HR and FAR effects that match recognition data (Figure 2.4).

This decision process, however, is not entirely satisfactory. The study lists here are mixed, containing words randomly sampled from the dataset, so the decision criterion needs to be adjusted on a per-stimulus basis. As discussed, this is necessary to achieve proper FAR trends. Attaining a qualitatively correct false alarm effect was our criterion for choosing a decision process here. However, explaining

FAR effects with a criterion shift remains controversial. Stretch and Wixted (1998) provide evidence that the strength-based but not the frequency-based recognition mirror effects depend on criterion shift. However, Miller and Wolford (1999) argued that a signal detection account of a simple false memory paradigm does support criterion shift as a mechanism for generating recognition false alarms. This was refuted (Roediger and McDermott, 1999; Wixted and Stretch, 2000) in part by asserting the compatibility of the other models with such an account (Wickens and Hirshman, 2000).

It remains that if WAS–FE is to demonstrate WF-dependent recognition differences, then it must employ a criterion shift. However, some recent cases are able to demonstrate that subjects modulate their decision criteria on-line according to stimulus class to optimize performance (Heit et al., 2003; Benjamin and Bawa, 2004). This strategic use of multiple criteria may be driven by self-knowledge of the category-dependent memorability differences of probe stimuli (Strack et al., 2005). The multiple-criterion decision process required by WAS–FE is in line with these observations. Note also that we intentionally constrained our decision-making process to be simple, plausible, and within the scope of WAS–FE.

Finally, criterion-independent performance is illustrated by the ROCs. For both small (Figure 2.3a) and large (Figure 2.3b) list sizes, the trial-averaged ROC for the bin of LF words has higher HRs and lower FARs than that of HF words. These characteristics correspond to the Λ_6–Φ and Λ_1–Φ comparisons, respectively. The shapes of these characteristic curves derive from the non-Gaussian form of the corresponding energy distributions (Figure 2.2a, right column). They largely resemble those of other recognition models except they are not symmetric around the negative diagonal. These ROC examples also demonstrate two performance effects of the number of trained stimuli. The low-load condition (Figure 2.3a)

shows high absolute performance, but a relatively small WFE; the high-load condition (Figure 2.3b), however, shows worse overall performance, but a larger difference between LF and HF words. These are capacity effects of the attractor-based FD mechanism. Larger stored lists entail higher synaptic load and reduced recognition accuracy. Further, we can infer that such capacity effects hurt the performance of HF words more than LF words. This WF-dependence may be a result of the sparse encoding of HF words in the WAS: weaker attractors are more sensitive to the perturbations of over-learning than the strong LF-word attractors. So, for a given reference pool size, increases in study-list length push the network closer to capacity, decreasing HRs and increasing FARs regardless of WF class. This means that WAS–FE exhibits a list-length mirror effect, which previous single-process models have also demonstrated (Shiffrin et al., 1990; Shiffrin and Steyvers, 1997; McClelland and Chappell, 1998). Conversely, for constant study-list length, this predicts better overall recognition performance and a smaller WFE for subjects with relatively less background experience (e.g., children vs. adults).

2.5.4 Role of contextual information in recognition

Dual-process recognition theories employ asymmetric recollective processing as the basis of the HR effect for LF words; the FAR effect is due to error-prone familiarity processing of similarly encoded HF words (Guttentag and Carroll, 1994, 1997; Reder et al., 2000; Arndt and Reder, 2002). This account is supported by evidence from pharmacological dissociation of recollection (Hirshman et al., 2002; Mintzner, 2003), but not to the ultimate exclusion of single-process accounts (Malmberg et al., 2004). Indeed, it seems that both familiarity and recollection are involved but the exact nature of their interaction is not yet definitively characterized (for review, see Yonelinas, 2002).

As a decision-making recognition model, WAS–FE is not purely a single process familiarity model. The process interaction implied here is different from the frequency tradeoff proposed in the source of activation confusion (SAC; Reder et al., 2000) model. In the decision comparison Λ_i–Φ, words from the study context are treated categorically as members of their respective WF class. However, non-study probes are not likewise differentiated. This is a contextual distinction that is necessitated by WAS–FE as discussed above, but this is not to say that some recollective process is making perfect old–new discriminations only to then involve an error-prone familiarity process. The contextual distinction consists of the subject having formed stimulus categories, such as frequency, only for recently studied stimuli. These categories then inform the decision process. Both IT cortex and PER are implicated in highly plastic category formation (Erickson et al., 2000; Miller, 2000; Miller et al., 2003), thus the formation of such WF categories of recent semantic stimuli is plausible. Also, further episodic information could allow discrimination between, for example, several study lists in a session. This could be modeled within the framework of WAS–FE by integrating a representation of a time-varying context signal (e.g., Howard and Kahana, 2002). Regardless, the main point here is that the decision comparison requires some contextual distinction of this sort in order for WAS–FE to yield a proper WFE mirror effect.

Dual-process models typically employ differential recruitment of recollective processing. Physiologically, this would be evident as WF-modulation of activity in regions such as the hippocampal formation and MTL. However, WAS–FE predicts that such areas, including PER, differentiate old–new responses but do not exhibit frequency dependence. It also predicts that the area responsible for semantic representation and processing shows WF-modulated activity. Using event-related fMRI at retrieval, de Zubicaray et al. (2005) sought to test predictions such as these

and found two main effects. First, recollection-specific MTL regions with significant old–new responses did not show WF modulation. Second, the LF word HR advantage was associated with left lateral temporal cortex (LTC) activation. Evidence suggests that LTC but not MTL structures are necessary for lexical-semantic information processing (Levy et al., 2004; de Zubicaray et al., 2005). Thus, LTC is well-positioned as a possible semantic input region for familarity processing in PER. de Zubicaray et al. (2005) suggest these results are consistent with context-noise models of the recognition WFE, but they are also consistent with our account here. Recently, EEG techniques have been able to dissociate verbal from nonverbal retrieval (Hwang et al., 2005), indicating the possibility of investigations using higher temporal-resolution methods. More such studies are needed to complement the large body of behavioral data.

2.5.5 Conclusion

In the present work, we take advantage of an empirically-determined model of semantic space to demonstrate a benchmark effect of human memory. Using the WAS as an input space for a Hopfield model of perirhinal familiarity processing, we found a word-frequency effect on familiarity distributions that can be explained as a function of the small-world structure of the semantic space. This structure, characterized by tight local clustering of rare words, implies that word frequency is non-intuitively encoded into the semantic structure of language. We argue that the model components plausibly capture the salient features, respectively, of semantic representation and neurobiological familiarity processing. Thus, we suggest that lexical-semantic structure forms a causal basis for the recognition WFE. Further, we show that a frequency-dependent criterion shift produces a WFE mirror effect without requiring log-likelihood computations to bring about old–

new symmetry. This entails a role for dual-process involvement in recognition contrary to previous models but consistent with some recent imaging data. Finally, we hope to have demonstrated the utility of relatively simple, but specific and salient, models of complex biological systems and likewise the importance of establishing an appropriate interpretative context.

Chapter 3

Hippocampal Network Dynamics Constrain Representations of Novel Environments

Everyone was staring at a television set hooked up to a development box for the Sony Playstation. There, on the screen, against a single-color background, was a black triangle.

"It's a black triangle," [our HR lady] said in an amused but sarcastic voice. One of the engine programmers tried to explain, but she shook her head and went back to her office.

Afterwards, we came to refer to certain types of accomplishments as "black triangles." These are important accomplishments that take a lot of effort to achieve, but upon completion you don't have much to show for it — only that more work can now proceed.

—Jay Barnson, The Rampant Coyote
http://www.rampantgames.com/blog/2004/10/black-triangle.html

Summary

The fast and contiguous association of entorhinal realignment and hippocampal global remapping constrains the time-course of initial code formation in hippocampus. As a real-time readout of grid-cell response structure, we demonstrate a minimal competitive rate model of homogeneous nonlinear place units with fixed synapses from a heterogeneous set of simulated grid maps. Competition is mediated by global recurrent inhibition. When the input excitation is balanced with the inhibitory gain, population spatial maps are rapidly available that exhibit both sparse activity and full spatial representation. The detailed competitive balance across the environment makes the place code highly input-sensitive. This sensitivity may contribute to pattern separation, so we use the model to explore remapping as a function of entorhinal realignment (Chapter 4). We suggest that balanced feedback competition on heterogeneous grids can provide a short-timescale prototype spatial representation as the starting point for subsequent encoding as novel environments become familiar with experience.

3.1 Introduction

Spatially modulated place cells in dorsal hippocampus (O'Keefe and Dostrovsky, 1971; O'Keefe, 1976; O'Keefe and Nadel, 1978) receive spatial information from grid cells in medial entorhinal cortex (MEC) (for review, see Section 1.3; Fyhn et al., 2004; Hafting et al., 2005; Witter, 2007). As such, the spatial component of hippocampal representations of an animal's local environment likely depends on the particular alignment of grids concurrently available in MEC. Critically, MEC grids can realign nearly instantaneously, a process characterized by randomization of spatial phases and contiguity with hippocampal global remapping (Fyhn

et al., 2007). The fast, simultaneous transitions observed in MEC and subregion CA3 suggest hard-wired network dynamics in hippocampus may be reading out cortical response changes (Colgin et al., 2008). In novel environments, or when confronted with substantial amounts of contextual change, the hippocampal spatial representation must reorganize quickly. While CA3 may be essential for detecting novel spatial information (Lee et al., 2005) and encoding those changes using various mechanisms of synaptic plasticity (Lee and Kesner, 2002; Nakazawa et al., 2003; Haberman et al., 2008), spatial learning based on synaptic reorganization can be a long-timescale process that requires non-trivial experience with the new conditions (Wilson and McNaughton, 1993; Gerstner and Abbott, 1997; Leutgeb et al., 2004; Karlsson and Frank, 2008). This is not commensurate with the speed of observed transitions under remapping conditions. Also, given the possible link between hippocampal spatial maps and episodic memory encoding (Knierim et al., 2006; Leutgeb et al., 2006; Leutgeb and Leutgeb, 2007), basic considerations of mnemonic requirements in novel environments suggest the importance of having a good approximation to the final map available immediately. Thus, rapid learning in unfamiliar contexts may depend on the short-timescale formation of operationally adequate, if initially unreliable, spatial representations.

Since plasticity is activity-dependent, the initial first-pass responses of a network of place cells have significant influence over the spatial representation that develops with further experience. If this initial activity state is a poor spatial representation, then substantial synaptic reorganization and additional familiarization would be required to produce an adequate spatial map. However, if a naïve network is able to provide an operationally adequate spatial representation, based only on the alignment of MEC grids, then mechanisms of plasticity would only need to refine the map (e.g., improve spatial specificity, sparsity, reliability). That

is, hard-wired activity could bootstrap the rapid formation of a spatial representation by enabling faster, more efficient learning processes. This would prevent the need for major reconfiguration through synaptic modification. Here, we explore recurrent inhibition, tightly coupled into a network of place units, as a putative mechanism for rapid spatial coding.

Inhibition is strong and prevalent throughout the hippocampus, and is critical to not just the stability of hippocampal activity but also its computational capabilities. Generally, interneurons sustain high firing rates and comprise just a minority of hippocampal cells, but they exert tremendous control over sparse principal cells (Buzsáki et al., 2007). Further, the functional coordination of dentate gyrus (DG) and CA3 is tied to both strong feedforward inhibition (Buckmaster and Schwartzkroin, 1995) and the significant targeting of recurrent interneurons over pyramidal cells by granule cell projections in the mossy fibers (Acsady et al., 1998). Interneurons in CA3 have separable pathways for synaptic transmission of recurrent and feedforward inputs, consisting of modifiable synapses (Pelletier and Lacaille, 2008). This suggests a critical role for recurrent inhibition in hippocampal computation. The activity of interneurons in CA1 decreases upon entering a novel environment (Wilson and McNaughton, 1993) but ramp back up as familiarity increases (Frank et al., 2004). In contrast, inhibitory activity in DG is enhanced upon introduction to a novel environment (Nitz and McNaughton, 2004). We suggest that these modulations reflect dynamic changes in recurrently driven activity that set the stage for the rapid learning of novel spatial information.

We present a minimal model of competitive network dynamics that can develop qualitatively relevant spatial representations of environments by integrating a set of heterogeneous MEC grids. The model creates sparse spatial codes over time by balancing afferent excitation from the cortical inputs with competition for

representation of the environment driven by global recurrent inhibition. We do not place constraints on the spatial metrics of grid-cell responses (Section 1.3) as has been done in other models (Solstad et al., 2006). Thus, creating the representation of an unfamiliar environment does not require any special configuration of the grids beyond random spatial phases and orientations. Further, the competitive balance is heterogeneous throughout the environment, depending on the detailed relative distribution of cortical inputs, and so contributes to pattern separation by amplifying small input and place coding changes (Lee et al., 2004; Leutgeb et al., 2007). This may dynamically enhance remapping despite local coherence in cortical response changes (Fyhn et al., 2007). In Chapter 4, we explore hippocampal remapping under various conditions of MEC realignment.

3.2 Model

3.2.1 Cortical inputs and connectivity

We consider a network model of 500 grid-cell response maps (Figure 3.1) projecting with random weights to 300 place units (Figure 3.2). The spacing and field size of the grids are determined by linear approximations of observed grid-cell spacing (Equation 3.2) and field-size (Equation 3.3) data, respectively. Unless otherwise specified, the grid maps have randomized orientations and spatial phases with spatial frequencies corresponding to the first 1mm of the dorsoventral extent of dorsocaudal MEC (Figure 1.5c). We use this model to derive spatial representations of a 1m square environment in order to match the grid scale on which the initially observed responses are based (see Methods and Hafting et al., 2005). New environments are created by sampling new random spatial phases and orienta-

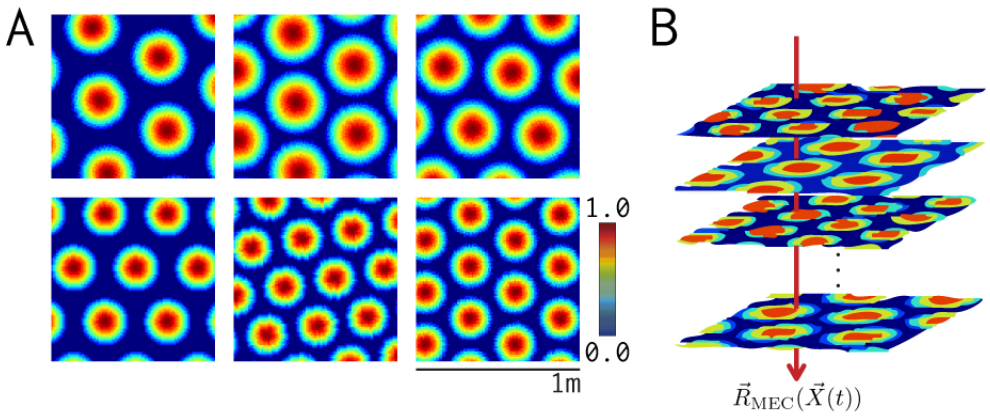

Figure 3.1 Example grid-cell response maps A. Grid maps have random spatial phases and orientations representing periodic coverage of a 1m square environment. B. Illustration of a population input vector being elicited by the simulated rat's location in the environment $\vec{X}$ at a given time t.

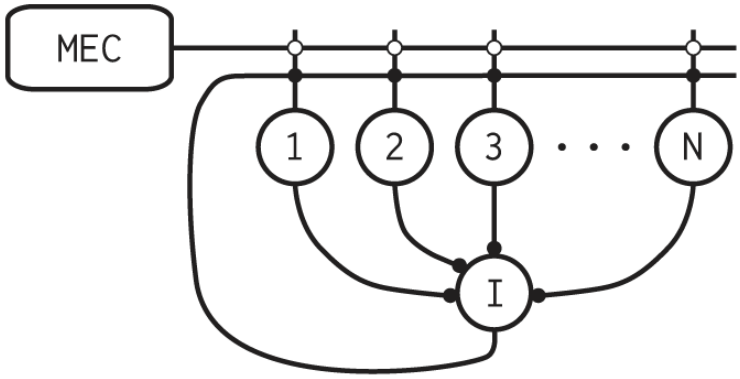

Figure 3.2 A minimal model for dynamically integrating MEC grid-cell responses. Schematic of the spatial map formation model presented here with equal-gain global inhibition enabling competition among place units.

tions for the grids in the input set and reconstructing the response maps.

Input convergence, or fan-in, is fixed in the weight matrix $\mathbf{W}$ so that each place unit is innervated by 50% (250 grids) of the input set. This respects previous estimates on the range of MEC inputs to dentate granule and hippocampal pyramidal cells (100–1000; Amaral et al., 1990). For each random network that we simulate, we set the afferent weights for each place unit to a random permutation of some

random reference weight distribution (see Section 3.3.3 for details). We do this so that all of the place units have not only the same overall synaptic gain, but identically distributed afferent weights. This means that any heterogeneity of the output responses cannot be the result of differences in the detailed connectivity patterns across the network of place units. Finally, when we refer to a "network–environment pair", we indicate that a random network (determined by a random weight matrix $\mathbf{W}$) was used to simulate the spatial map for the cortical representation of a random environment (determined by random spatial phase and orientation vectors).

3.2.2 Network interaction

Network interaction consists solely of recurrent inhibitory feedback mediated by global recurrent inhibition (Figure 3.2), which we represent as an instantaneous linear readout of fluctuations in the activity of the place-unit population. Each place unit receives inhibitory input at gain J_{inh} and so has a recurrent neural field $h_{\text{rec}} = -J_{\text{inh}} \langle \vec{R} \rangle$ at any time-step in the simulation, where $\langle \cdot \rangle$ indicates a population average and $\vec{R}$ is the population firing rate vector for the place units. The afferent field is the MEC input current $\vec{h}_{\text{aff}} = \mathbf{W}\vec{R}_{\text{MEC}}$, where $\vec{R}_{\text{MEC}}$ is the population firing rate vector for the grid inputs (Figure 3.1b). Each place unit integrates the total field $\vec{h}$ (see Methods and Equation 3.4) within a nonlinearity characterized by a half-rectified hyperbolic tangent:

$$\tau \frac{\mathrm{d}\vec{R}}{\mathrm{d}t} = -\vec{R} + \left[\tanh\left(\frac{\vec{h} - \lambda}{\sigma} \right) \right]_+ \tag{3.1}$$

where τ is the integration time constant of the place units. The smoothness of the nonlinearity is fixed at $\sigma = 0.1$, but the position of the nonlinearity with respect to the total field $\vec{h}$, set by λ, regulates the amount of afferent excitation that drives the

network. The inhibitory gain J_{inh} and the position of the nonlinearity λ are the two primary parameters of the dynamics of this model. J_{inh} determines the strength of the recurrent competition while λ modulates the afferent excitation from the simulated grid maps. Since the recurrent interaction, the nonlinearity, and the weight distributions are all homogeneous across the network, any spatial heterogeneity within the network response follows from the structure of the cortical inputs and the hard-wired network dynamics described here.

3.3 Methods

3.3.1 Software and graphics

The model simulation and analysis software were developed as custom Python packages (http://www.python.org/). These packages use the NumPy library for its ndarray implementation of numerical arrays (http://numpy.scipy.org/) and the SciPy library for some scientific computing functionality (http://www.scipy.org/SciPy/). Plots were created in IPython interactive sessions (http://ipython.scipy.org/) using the matplotlib plotting library (http://matplotlib.sourceforge.net/) and imported into Adobe® Illustrator® CS4 (http://adobe.com/Illustrator/) for final composition. Two-dimensional matrix arrays were converted into RGB image data using matplotlib colormaps and saved directly in the lossless Portable Network Graphics format (PNG; http://www.libpng.org/pub/png/) using the Python Imaging Library (PIL; http://www.pythonware.com/products/pil/).

For both simulations and data analysis, all maps of the 1m square environment were represented as 100×100 matrix arrays, so that each pixel represents 1 square cm of the environment. Population maps, for both the simulated MEC and place

network, are represented as 3-index arrays with the unit index along the first dimension and the spatial indices along the last two (e.g., Figure 3.1).

3.3.2 Simulated MEC grid-cell responses

The simulated MEC grid-unit responses are modeled phenemonologically. The grid metrics are computed based on the linear regressions between grid spacing and recording distance ventral of the postrhinal (POR) border and between field size and grid spacing (Supplemental Figure 4e–g of Hafting et al., 2005). For a random population vector of recording locations $\vec{d}$, uniform over the range 0–1 mm from the POR border, the corresponding grid spacing vector

$$\overrightarrow{\text{spacing}}\ (\text{cm}) \ = 30 \ + \ 20\,\vec{d} \ + \ \vec{\epsilon}_{\text{s}} \tag{3.2}$$

and field radius vector

$$\overrightarrow{\text{radius}}\ (\text{cm}) \ = \frac{0.57}{\sqrt{\pi}}\overrightarrow{\text{spacing}} \ + \ \vec{\epsilon}_{\text{r}} \tag{3.3}$$

are computed, where $\vec{\epsilon}_{\text{x}}$ are uniform random vectors on $\{-1.5, 1.5\}$ to approximate the observed variance of the grid metrics. Extended response maps of 211×211 pixels (about 4.5m square) are computed for each grid unit by placing Gaussian firing fields (half-max to peak of a two-dimensional Gaussian function) at the vertices of a triangular grid with random spatial phase and 0 degrees orientation. All fields have the same peak rate of 1.0 and are homogeneous within a given grid response map. These extended maps serve as the basis of grid-unit response maps. Before each simulation in a new environment, the response maps for each grid unit are computed directly as translations and rotations, as defined by spatial phase and orientation vectors, of a 100×100 pixel window superposed on its corresponding extended map. The cortical input vector, $\vec{R}_{\text{MEC}}$, at any location in the environ-

ment is then available by indexing the MEC population map consisting of these environment-specific response maps.

3.3.3 Spatial map simulations

Having the pre-computed MEC responses, we can then set up the model simulation to create a spatial map. An afferent weight matrix $\mathbf{W}$ is created or restored from a previous simulation. To create a random $\mathbf{W}$, a reference weight distribution of uniform random weights on $\mathcal{U}\{0, 1\}$ from a random subset of 250 grid units is sampled (i.e., 50% fan-in connectivity from the set of 500 simulated grid units used here). Each row of $\mathbf{W}$, corresponding to a place unit's afferent weights, is set to a random permutation of this reference distribution. Next, the population rate vector $\vec{R}$ is initialized to the zero vector and a trajectory through the environment is constructed. For all spatial map simulations presented here, a checker-pattern raster-scan trajectory is used that samples every other pixel in the environment. We did this for computational efficiency, as the spatial scale of place-unit responses is significantly larger than a single pixel (1 square cm).

The simulation then consists of sequentially clamping the network to the input vectors $\vec{R}_{\text{MEC}}$ corresponding to the scan locations in the trajectory. Throughout, the rate vector $\vec{R}$ is evolved according to Equation 3.1 using fourth-order Runge-Kutta numerical integration. We integrate at a resolution of $\Delta\, t = \tau/10$, so that a simulation interval of duration τ comprises 10 discrete time-steps. For each time-step, the total neural field $\vec{h}$ is computed, consisting of the afferent and recurrent fields. In practice, the two component fields are multiplied by constants that bring them approximately to unity order:

$$\vec{h} \;=\; \alpha\, \vec{h}_{\text{aff}} \;+\; \beta\, \vec{h}_{\text{rec}} \;=\; \alpha\, \mathbf{W}\, \vec{R}_{\text{MEC}} \;-\; \beta\, J_{\text{inh}}\, \langle \vec{R} \rangle \tag{3.4}$$

where $\alpha = 0.05$ and $\beta = 100$. This allows the parameters J_{inh} and λ to be approximately unity order. The clamp durations used in simulations here are typically 5–6 τ. The first pixel to be clamped is held for $10\,\tau$ to allow network activity to come online. At the end of each input clamp, the final population rate vector associated with that particular pixel is stored. The rate vector is not reset again except in the simulations shown in Figures 3.20–3.26, in which $\vec{R}$ is set to the zero vector after every pixel in order to synchronize the onset of the network response across the environment.

3.3.4 Characterizing the spatial output

Rate response maps are constructed from the stored rate vectors resulting from the rastern-scan of the environment. The sampled pixels are set directly in a population array based on the stored data, while the non-sampled pixel responses are computed as averages of all adjacent sampled pixels. Each response map is then median-filtered with a 3×3 kernel to remove single-pixel artifacts due to the averaging process.

Spatial activity and place fields are then determined by several activity thresholds. We consider noise to be all activity below 10% of the population maximum rate or below 15% of a place unit's own maximum rate. Then, based on these thresholds, we determine and store all contiguously active regions of more than 50 square cm as active place fields. Using binarized maps demarcating active fields, then, we compute place-unit coverage maps, population coverage maps and population representation maps of the environment. Various population spatial map, place unit, and place field statistics are then computed based on the coverage, representation and rate response maps. Place units with no active fields are considered "dead units" for the environment and are not included in place unit statistics.

Spatial map sparsity is determined as the overall proportion of dead units in the population within a given environment.

3.3.5 Parametric analysis and visualization

The two-dimensional parameter sweeps (Figures 3.3 and 3.4) are the result of simulating a single random network with a single random environment for a 15×15 grid of regularly–spaced points in (J_{inh}, λ)–space. Spatial coding charateristics were stored for each sample point and a 256×256 matrix array was created using bilinear interpolation over the sampled data for visualization of the results over the extent of the sampled parameter region. The one-dimensional parameter sweeps (Figure 3.12) use a single random environment for each parameter condition shown; at each point in the sweep, 10 random networks are simulated to produce the resulting spatial map data for which means and standard errors of the mean (SEMs) are shown.

To visualize the sparsity of spatial codes and the quality of code transitions across the environment, we computed pair-wise population rate correlations of a diagonal traversal through the environment (Figures 3.7a and 3.8). For each of the 100 pixels from $(0, 0)$ (lower left corner) to $(100, 100)$ (upper right corner), we computed the correlation between every pair of population rate vectors in the map:

$$\mathrm{C}_{ij} = \mathrm{corr}\left(\vec{R}(i,i), \vec{R}(j,j)\right) \text{ for } i,j \in \{0, 100\}, \tag{3.5}$$

where $\vec{R}$ is a population rate vector defined across the environment, i and j are spatial indices in pixels (cm), and corr$(\cdot)$ is the Pearson correlation function. The width of the diagonal band of the resultant 100×100 correlation matrix indicates the scale of spatial correlations.

To visualize changes in rate distributions across λ (Figure 3.16) and within on-

set dynamics (Figure 3.23), we constructed smoothed estimated probability density functions (PDFs). This method was adapted from the methods of Karlsson and Frank (2008). For a given rate distribution, we computed a fine-grained cumulative histogram (1000 bins for data of size 300), extended its end-points to reduce boundary effects, and convolved it with a Gaussian kernel for smoothing. For the smoothed PDF, we computed the differential of the smoothed cumulative data, cropped it to the original data range, and normalized the resulting densities to their trapezoidal integral. The standard deviations (SDs) of the smoothing kernels used are 0.1 (Figures 3.16 and 3.23a) or 5% of the data range (Figures 3.10b, 3.14c, 3.19 and 3.23b). In all cases, the width of the kernel was 10 times the SD.

To visualize and compare population fluctuations across the environment, we computed per-pixel rate vector norms for both the simulated MEC inputs and spatial map outputs. To compute the norms, we squared the elements of a population map, summed along the first dimension (the population index), and took the square root of the elements of the matrix array containing the resulting environment map:

$$\left|\vec{R}(x,y)\right| = \sqrt{\sum_{i=1}^{N}\left(\vec{R}_i(x,y)\right)^2}, \tag{3.6}$$

where $\vec{R}$ is a population rate vector defined for all points (x, y) across the environment, and N is the size of the population.

Population autocorrelograms were computed to visualize the spatial scale and correlational structure of the MEC inputs and spatial map outputs (Figure 3.18, top row). We computed these by Fourier domain multiplication of a population map with its complex conjugate, followed by summation along the first dimension (the population index) and an inverse Fourier transform:

$$\text{AutoCorr}(\vec{R}) = \mathcal{F}^{-1}\left(\sum_{i=1}^{N}\left[\mathcal{F}(\vec{R}_i) * \mathcal{F}(\vec{R}_i^*)\right]\right) \tag{3.7}$$

where $\vec{R}$ is a population ratemap defined across the environment, $\vec{R}^*$ is its complex conjugate, and $\mathcal{F}$ is a discrete Fourier transform across the spatial dimensions. Linearizations of the autocorrelograms (Figure 3.18, bottom row) were computed with a radial average, discretized into 1 cm bins, relative to the central peak. To show the full symmetric autocorrelation, the radial average was then copied and reflected across the mantissa.

3.4 Results

We begin here by using parameter sweeps to assess the dependence of various spatial map properties on the primary dynamic parameters of the model. One of the properties we consider, the total coverage of the environment by active place fields, allows us to use this model to investigate global remapping in Chapter 4. These parameter sweeps allow us to determine a parameter reference-point where informative spatial representations are produced. We proceed to show that this reference-point coincides with an excitatory–inhibitory balance state in the network and that this balanced spatial coding allows effective competition for place representation. We then present distributions and statistics of spatial map properties for random sample sets. Using autocorrelations, we show that spurious place fields tend to be the result of the periodic grid structure of the cortical inputs. We then assess the onset dynamics of spatial coding and show that the rate code is slower to converge than the place code, which is available almost immediately.

3.4.1 Parameter dependence of spatial coding

First, we examine the interaction between the inhibitory gain J_{inh} and the nonlinearity threshold λ by sweeping ranges of both parameters for a single network–

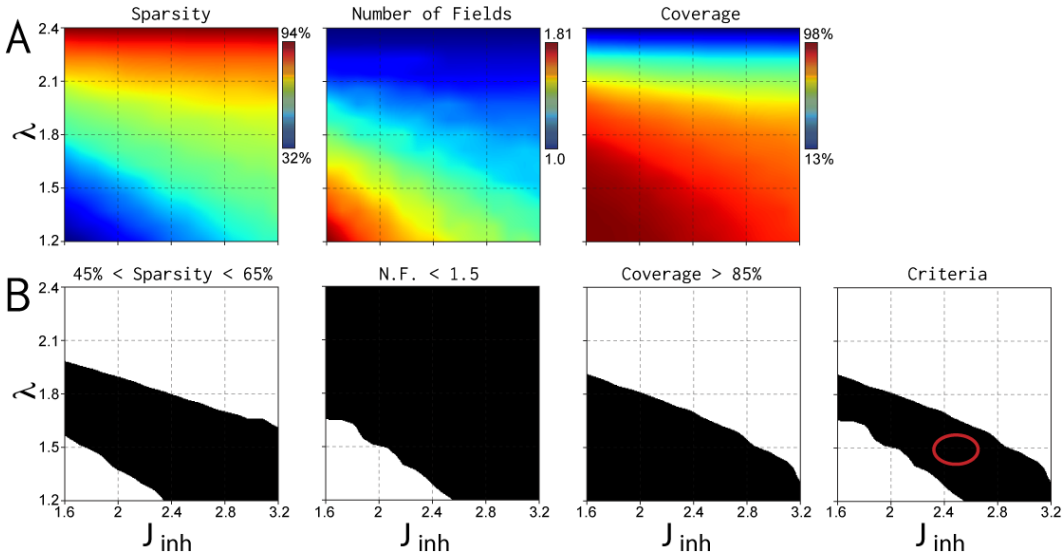

Figure 3.3 Searching the $(J_{\text{inh}}, \lambda)$ parameter space for sparse but representational spatial coding. A. Parameter sweeps for a network integrating an environment (data is linearly interpolated from 15×15 parameter points). Several key criteria are shown: the proportion of dead units in the network (left); the average number of place fields for active units (middle); and percentage of the environment overlapped by at least one active place field (right). B. Conditions on spatial map criteria for qualitatively relevant spatial coding (black indicates condition met). The intersection of these conditions (right) yields a band in $(J_{\text{inh}}, \lambda)$–space. Red ellipse indicates the region around the reference parameter point we will use throughout.

environment pair. We simulated a parameter sweep (see Methods) of J_{inh} from 1.6 to 3.2 and λ from 1.2 to 2.4. Both parameters independently and additively increase network sparsity and decrease the average number of place fields per active unit (N.F.) and the proportion of the environment overlapped by active fields (coverage) (Figure 3.3a). Sparsity varies from 32 to 94% across the sampled range, N.F. from 1.00 (indicating no secondary place fields) to 1.81, and coverage from 13 to 98%. These measures are strongly correlated across the samples: N.F.–coverage ($r = 0.88$); N.F.–sparsity ($r = -0.99$); sparsity–coverage ($r = -0.94$), where r is the per-pixel Pearson coefficient of the interpolated sweeps. Average spatial representation, also highly correlated to N.F. ($r = 0.99$), decreased from 3.6 fields per

pixel to just 0.14. These measures and others, such as place-field peak rate and area, vary smoothly across this parameter range (Figure 3.3). Though these spatial coding measures are highly and mutually correlated, they are significantly less correlated with rate-based measures (Figure 3.5). The high degree of inter-correlation is not surprising as these are different measures of the same underlying quantity: the amount of output activity. Notably, for all of these measures, as λ increases, the rate of change with J_{inh} tends to decrease. The position of the nonlinearity serves to threshold the amount of activity that can drive the recurrent interaction, limiting the effects of J_{inh}. Beyond $\lambda \simeq 2.4$, increasing λ shuts down all network activity.

To achieve biologically relevant spatial coding, we restrict spatial map sparsity to 45–65% and N.F. to no more than 1.5. We further impose the basic representational constraint that active place fields must cover at least 85% of the environment. While coverage of the environment is dependent on network size (larger networks would yield more place fields), we impose this constraint given our fixed network size and the desire to explore the complete remapping of one spatial representation into another (Chapter 4). These constraints demarcate different regions of the parameter range and their intersection yields a band in (J_{inh}, λ)–space (Figure 3.3b). We will consider a point within this band, where $J_{\text{inh}} = 2.5$ and $\lambda = 1.5$ (Figure 3.3b, red ellipse), to be a reference point for good spatial coding. Using the middle of the constraint intersection allows us to avoid the extremes of either recurrence- or threshold-dominated network dynamics.

Aside from these critical measures, there are other ways to quantify the output activity of the model. For the same parameter sweep spatial map simulations (Figure 3.3), we compute several population-based, unit-based, and field-based measures. Representation varies across the sampled range from 0.14 to 3.60 active fields per pixel and averages ~ 1.8 within the critical band (Figure 3.4a, left).

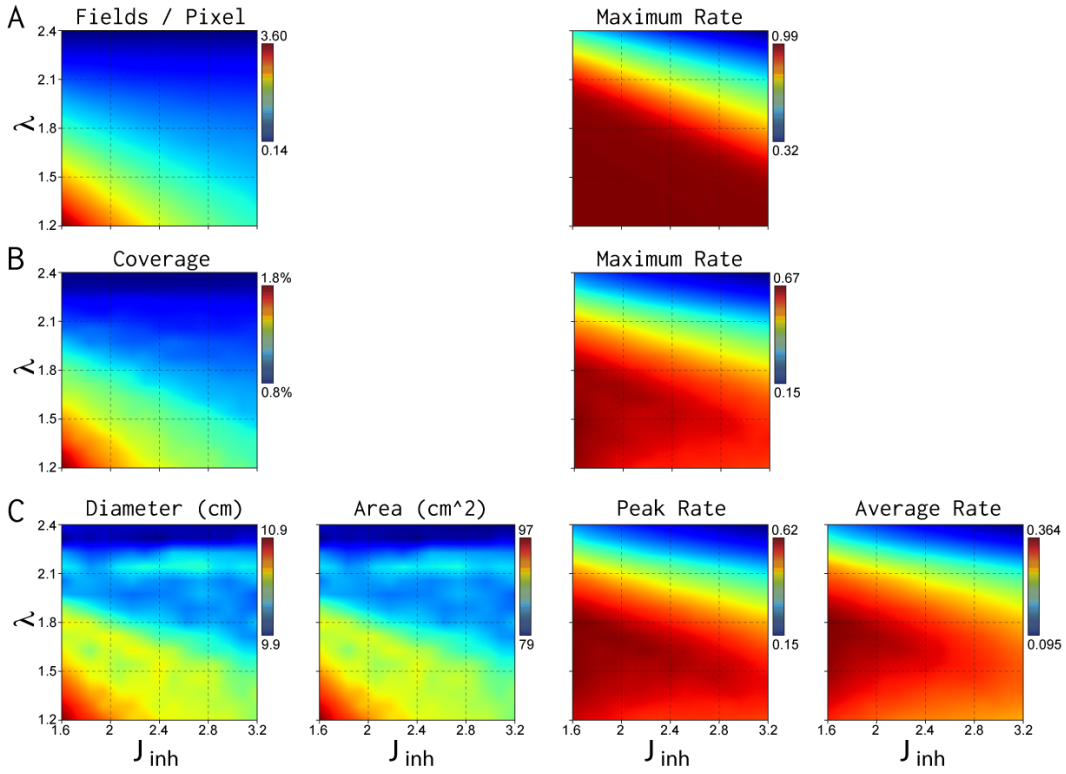

Figure 3.4 Parameter sweeps across J_{inh} and λ for a variety of spatial map characteristics. Additional measures are shown for the same set of simulated spatial maps as in Figure 3.3a. A. Population spatial map measures. Overall representation is measured as the average number of fields overlapping a given pixel (left), and the maximum rate attained in the population spatial map (right). B. Unit-based measures. Average active coverage by all fields (left) and maximum unit rate (right). C. Place field characteristics. Average field diameter is computed directly from field area assuming circularity. Average rate indicates the average firing rate throughout the extent of a place field.

As a measure of the spatial extent of place fields, its parameter-dependence is correlated with both average per-unit coverage (Figure 3.4b, left) and average place-field area and diameter (Figure 3.4c, left). However, the place field statistics exhibit decreased modulation with parameter changes: average diameters, which are computed directly from the areas, only vary from 9.9 to 10.9 cm across the range. This corresponds to field areas of 79–97 square cm. These place fields are generally smaller than those observed in even the most dorsal region of hippocampus.

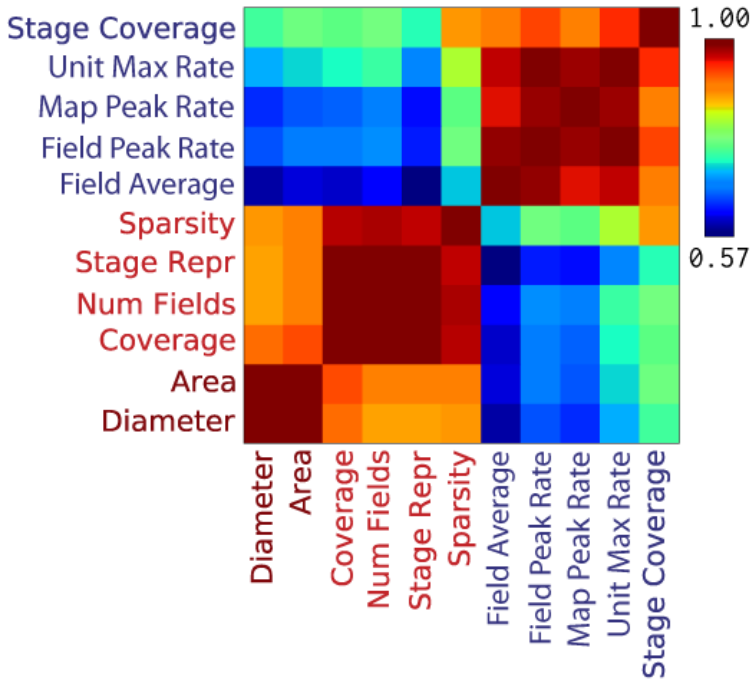

Figure 3.5 The parameter dependence of various quantifications of spatial map activity are strongly correlated. Spatial measures and rate-based measures are more correlated within groups than between groups. A correlation matrix of r^2 values is calculated from pixel-by-pixel correlations of parameter sweep data for any given pair of measures. These correlations constrain the model in order for various constraints on spatial coding for biological relevance to coincide (Figure 3.3b). For example, there is a minimally sufficient network size for enabling effective competition, which is the reason for the number of place units used here ($N = 300$).

However, associative synaptic modification and excitatory recurrence, which are not accounted for here, can play a significant role in increasing field size (see Discussion). We can also examine several rate-based measures. The maximum firing rate of the population (Figure 3.4a, right) is close to 1.0 for much of the parameter range, for $\lambda < 2.0$ when $J_{\text{inh}} = 0.0$ and $\lambda < 1.5$ when $J_{\text{inh}} = 3.2$, including the critical band (Figure 3.3b, right). This means that network output is at or near saturation of the possible dynamic range throughout the parameter region around the reference point. Per-unit and per-field averages of peak rates (Figure 3.4b–c, right)

vary with the population maximum except for non-monotonicity with respect to λ below $\lambda \simeq 1.5$ (Section 3.4.5). Since there appears to be differential parameter dependence between measures based on the spatial extent of activity and those based on firing rate, we can construct a correlation matrix for each of these measures (Figure 3.5). This matrix shows the Pearson r^2 values for pixel-by-pixel correlations of parameter sweeps of pairs of spatial map properties. Though the minimum r^2 here is 0.57, indicating substantial correlation between all measures, the division between rate and extent measures is clear. Interestingly, stage coverage (Figure 3.3a, right) is moderately correlated with both sets of measures. This indicates that population coverage may be integrative of the area–rate relationship for individual place fields (Figure 3.22a).

3.4.2 Inhibitory–excitatory balance for sparse codes

Fixing the nonlinearity at $\lambda = 1.5$, we can assess the effects of J_{inh} on the spatial specificity of individual place-unit response maps. For the case of $J_{\text{inh}} = 0.0$, or no recurrent inhibition, all place units respond at or near saturation throughout the environment (data not shown). This indicates that λ is significantly low relative to the afferent input distributions. Enabling recurrent inhibition at even a very low gain, $J_{\text{inh}} = 0.01$, deactivates roughly half of the response field for most of the network (Figure 3.6a, left). The improved specificity reveals the periodic structure of the underlying cortical inputs, but the responses are not place-like. Increasing J_{inh} to 0.1 silences most of the activity not due to local input peaks, so responses are characterized by multiple fields (Figure 3.6a, middle). Another order of magnitude increase in the inhibitory gain, to $J_{\text{inh}} = 2.5$, brings us to the reference point for spatial coding (Figure 3.3b, right). Here, the individual response are largely deactivated with a small number of heterogeneous regions of palce-like activity

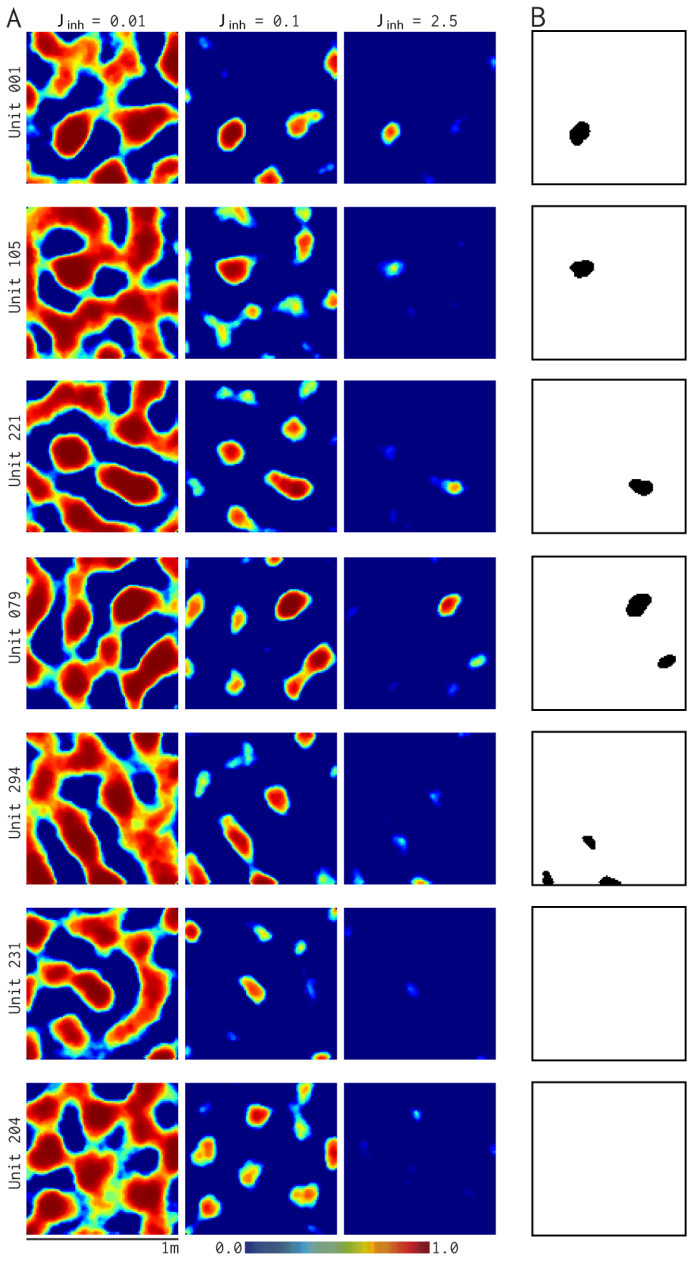

Figure 3.6 Inhibitory gain effects on the spatial specificity of place-unit responses with a moderate threshold ($\lambda = 1.5$). A. Example place-unit response maps of the same network–environment pair for $J_{\text{inh}} = 0.01$ (left), $J_{\text{inh}} = 0.1$ (middle), and the reference parameterization $J_{\text{inh}} = 2.5$ (right). B. Coverage maps demarcating active fields for the $J_{\text{inh}} = 2.5$ responses. The top three units have single-field responses while the next two show multi-field responses. The two bottom units are silent.

(Figure 3.6a, right). Applying noise-floor criteria to establish active place fields (see Methods), there is a diversity of single-field, multiple-field, and dead-unit responses across the population (Figure 3.6b). This diversity, sparsity, and irregularity of response is qualitatively indicative of hippocampal place-like activity. (Note that the same example network–environment pair presented here (Figure 3.6) is used for spatial map data presented in several figures to follow: Figures 3.7, 3.8, 3.9, 3.10, 3.11, 3.13, 3.14 and 3.18.)

Fixing the strength of recurrent inhibition at $J_{\text{inh}} = 2.5$, we can assess the effects of λ on the quality of spatial representation. Comparing the three cases of $\lambda = 1, 1.5,$ and 2, shows that linear changes in λ smoothly modulate the sparsity and coverage of the resultant spatial maps (Figure 3.7). The spatial scale of correlations and smoothness between code transitions is evident in per-pixel population rate correlation matrices for a diagonal traversal of the environment from $(0, 0)$ to $(100, 100)$ cm (Figure 3.7a). Increasing λ decreases both the off-diagonal correlations (due to secondary fields) and the spatial scale of correlation. Compared to the spatial outputs, the spatial scale of the MEC inputs is more consistent across the environment and off-diagonal correlations are more reflective of their periodicity (Figure 3.8). Representation maps, superpositions of binarized active fields that indicate the number of fields overlapping a given pixel, show both decreasing representation and coverage of the environment (Figure 3.7b). However, despite the decreasing coverage, the active field coverage remains evenly distributed. This distribution is better visualized by place-field center-of-mass (COM) locations (Figure 3.7c). For $\lambda = 1$, the map is densely populated with a substantial proportion of weak fields (low peak rates are indicated by dark colors in Figure 3.7c), but most of these weak fields are not apparent at the reference point $\lambda = 1.5$. Beyond $\lambda = 2$, the peak rates of the smaller set of remaining fields become suppressed as

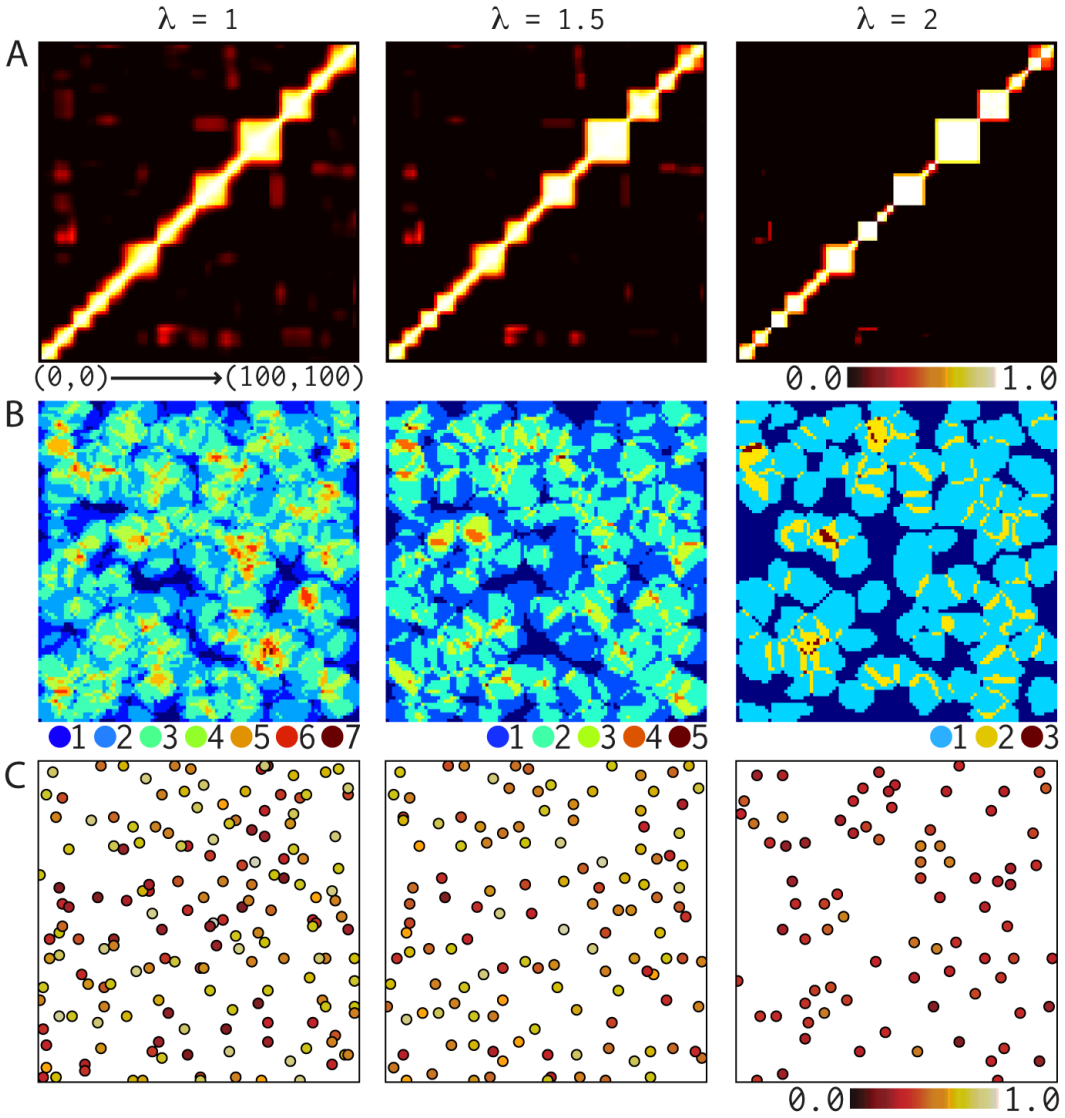

Figure 3.7 Threshold effects on spatial coding with moderate recurrent inhibition ($J_{\text{inh}} = 2.5$). **A–C** Simulations at four parameter points are shown: $\lambda = 1$, $\lambda = 1.5$ and $\lambda = 2$ with $J_{\text{inh}} = 2.5$. A. Correlation matrix of population rate vectors along a diagonal traversal of the environment from $(0, 0)$ to $(100, 100)$ cm. B. Environment representation is shown as a superposition of binarized place fields. Each pixel is colored by the number of fields that overlap it. C. Scatter plots of the place field centers-of-mass, colored by peak firing rate. With recurrent inhibition, λ modulates the degree of spatial representation, but place fields evenly cover the environment.

more and more afferent excitation becomes subthreshold and unable to drive the place population.

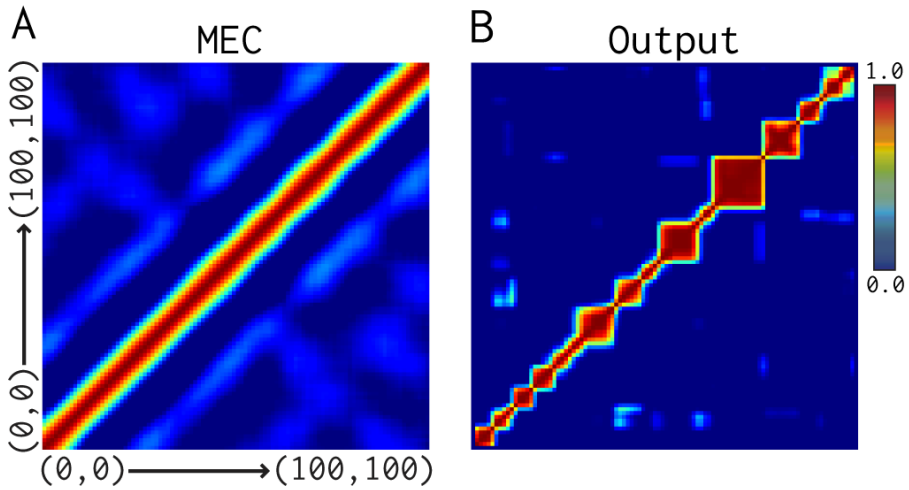

Figure 3.8 Population rate vector correlations for the simulated MEC inputs and balanced spatial outputs. Pair-wise correlations for a linear traversal of the environment from the lower left corner $(0, 0)$ cm to the upper right $(100, 100)$ cm. A. MEC correlation matrix, same cortical representation used as input for the simulations in Figure 3.7. B. Balanced spatial map output correlation matrix, same data as $\lambda = 1.5$ in Figure 3.7a.

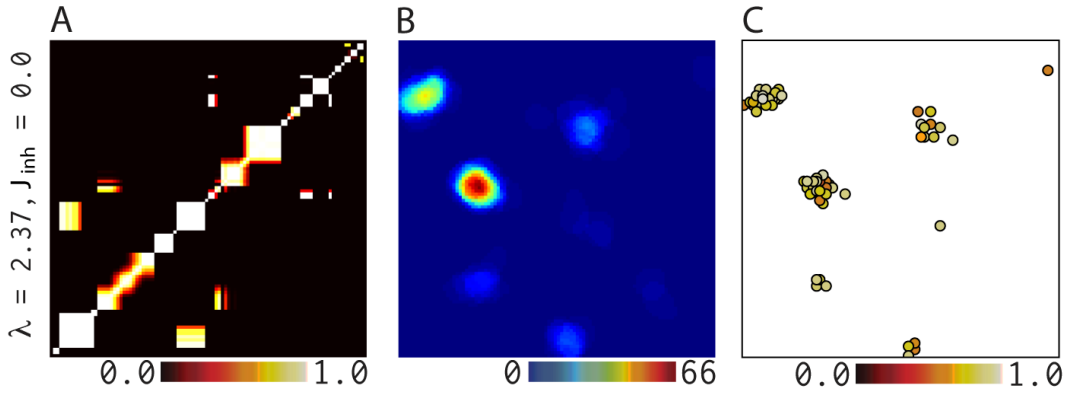

Figure 3.9 Threshold-based sparsity yields poor environmental representation. Using the same network as in Figure 3.7, the network sparsity of ($J_{\text{inh}} = 2.5, \lambda = 1.5$) is matched when recurrent is disabled ($J_{\text{inh}} = 0.0$) by increasing λ to 2.37. Without recurrent inhibition, spatial coding follows the strongest input peaks (Figure 3.10) so that the place fields cluster into correlated subgroups.

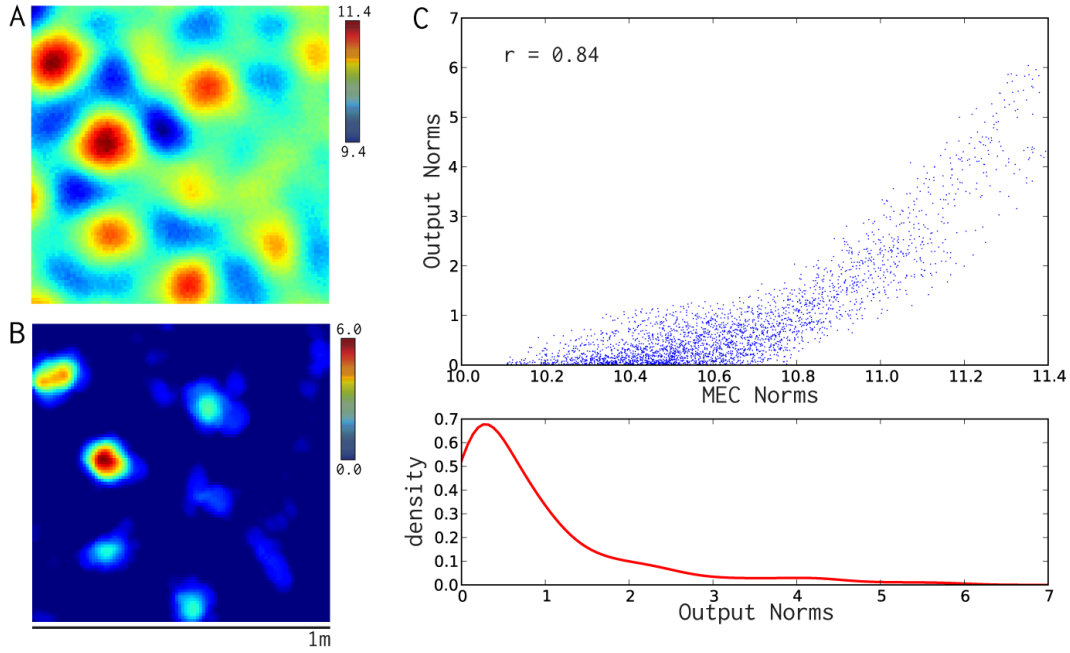

Figure 3.10 Threshold maps track fluctuations of the input across the environment. A. Population rate vector norms of the simulated MEC inputs set. B. Vector norms of the sparsity-matching threshold spatial map, formed without recurrent inhibition. C. Per-pixel scatter plot of the MEC vector norms and the output norms (top) shows strong correlation of input magnitude with output magnitude. Threshold-based sparsity yields outputs corresponding to the strongest cortical input peaks. The distribution of output norms (bottom) is highly skewed.

3.4.3 Feedback competition and informative representations

Since a threshold by itself is able to create individual place fields (McNaughton et al., 2006; Solstad et al., 2006), it is instructive to see the effects on spatial representation of relying on λ for output sparsity. Our example reference-point spatial map (Figure 3.7, $\lambda = 1.5$) has 128 active units out of 300 or, equivalently, a network sparsity of 57.3%. Turning off the recurrent interaction so that $J_{\text{inh}} = 0.0$, we have to increase λ to 2.37 to achieve the same network sparsity. For the linear traversal, off-diagonal correlations are stronger than in the recurrent cases (Figure 3.9a). More importantly, however, the place fields have clustered into a small number of

groups at the vertices of a quasi-triangular grid in the environment, resulting in severely irregular spatial representation (Figure 3.9b,c) that covers only 19.4% of the environment. The apparent grid corresponds to peaks in the magnitude of the MEC population rate vectors (measured as vector norms; Equation 3.6) across the environment (Figure 3.10a). The output population vector norms (Figure 3.10b), closely corresponding to the representation map (Figure 3.9b), are distributed significantly skewed to zero (Figure 3.10c, bottom). Those output norms that are significantly non-zero, however, closely follow the underlying cortical input peak. In a pixel-by-pixel scatter plot of the input–output transformation, the relationship is robust and supralinear (Figure 3.10c, top). Even so, the output norms for this case are highly correlated to the input fluctuations (per-pixel Pearson $r = 0.84$, $p \sim 0$). Without recurrent inhibition, the sparsity-matching threshold develops place fields at peaks corresponding to coincident vertices of the most correlated subset of MEC inputs.

To compare representational sparsity and the degree of information in the spatial code, we can assess the redundancy of place fields. A rate-independent method for doing this is to compute pixel overlap, equivalent to the area in square cm mutually represented by two place fields. If an environment is evenly represented, then fields across the environment will only overlap to a small degree with their closest neighboring fields. To create a similarity matrix, then, we sorted the primary place fields of active units by quadrants and computed pair-wise pixel overlaps with all other primary fields. For the reference-point spatial map, field similarity is sparse, uniform, and largely restricted to the local quadrant (Figure 3.11a). For the threshold-based spatial map, field similarity is strong throughout each quadrant and irregularly distributed so that one quadrant contains the vast majority of fields (Figure 3.11b). This results directly from the uniform distribution of

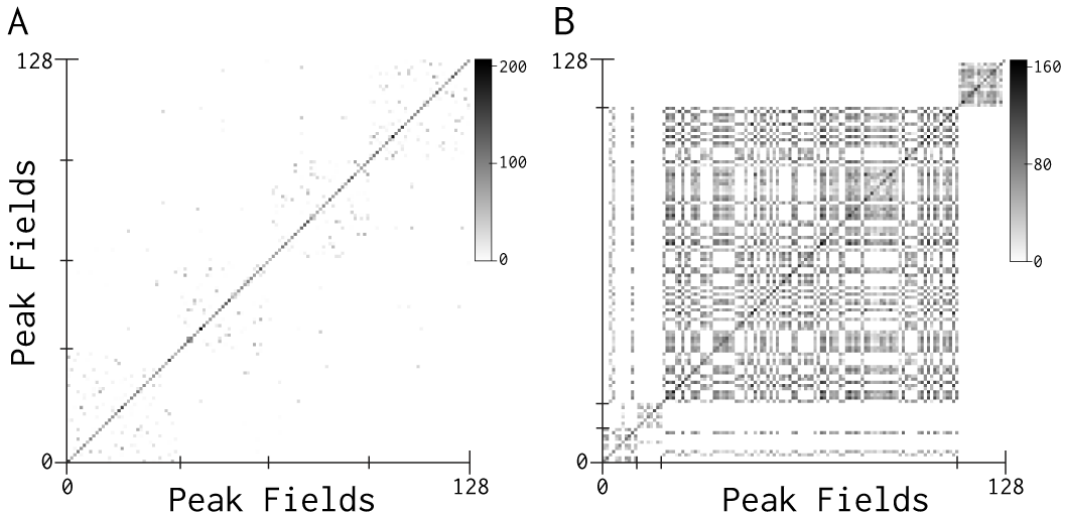

Figure 3.11 Place field similarity measured as pixel overlap. Place fields are partitioned into quadrants of the environment, depending on COM location, and pair-wise pixel overlaps are shown. A. Balanced reference-point spatial map. B. Sparsity-matching threshold spatial map. Place fields resulting from balanced spatial coding form a sparse, informative representation, whereas those from the threshold case are highly redundant with large blocks of correlated fields.

place fields in the former (Figure 3.7c) and the high degree of field clustering in the latter (Figure 3.9c)

To investigate the statistics of parameter dependence around the reference point, we computed sample sets of random networks forming representations of the same environment across various ranges of λ (Figure 3.12a) and J_{inh} (Figure 3.12b). Network sparsity and spatial specificity (N.F.) are shown along with the average place-field peak rate, which together give an indication of the quality of spatial coding. Notably, sparsity has a high gain around 50% with recurrent inhibition disabled (Figure 3.12a, top left). That is, threshold-based spatial coding is highly sensitive to small parameter changes. Also, the λ-range for relevant sparsity (50–60%) does not coincide with the λ-range for relevant spatial specificity (N.F. < 1.5) (Figure 3.12a, top). Further, network output progresses from saturated activity at $\lambda = 2.2$ to complete silence at $\lambda = 2.6$. So not only does it require fine tun-

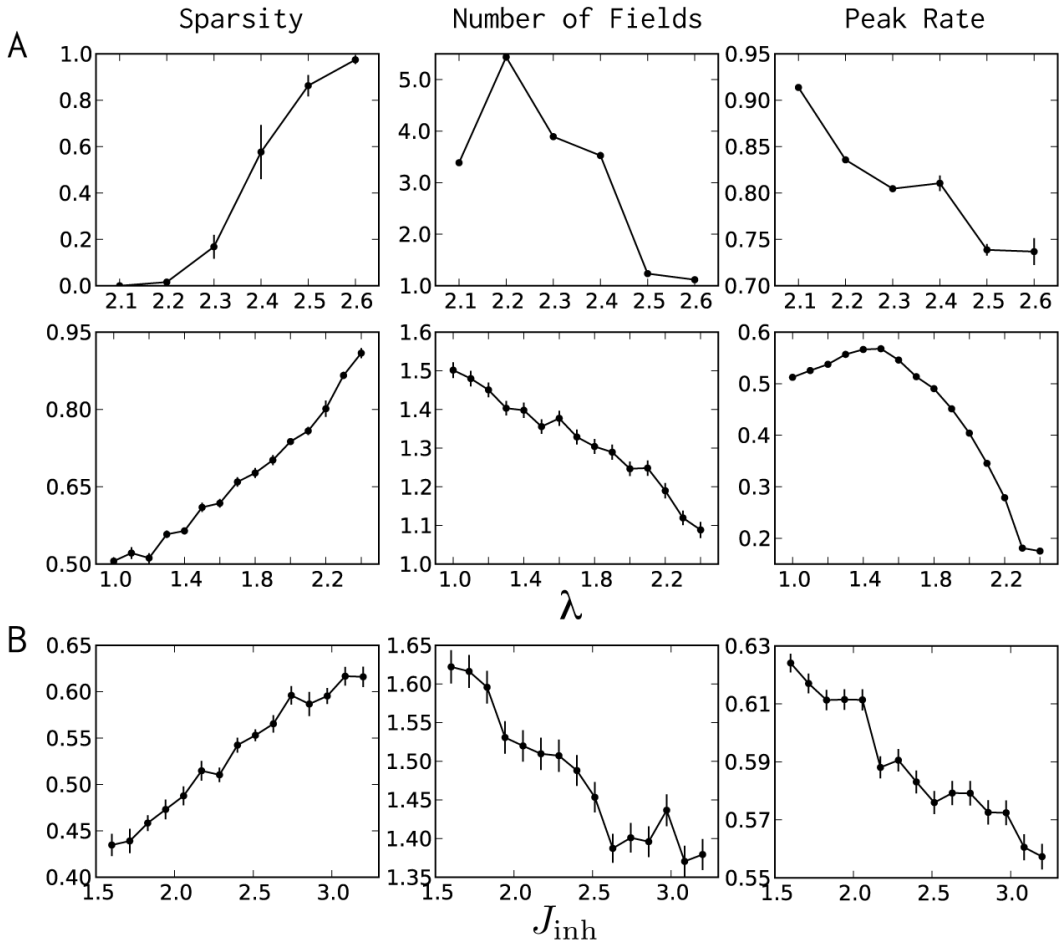

Figure 3.12 Statistics of spatial map characteristics around the J_{inh}–λ balance point. A,B Data shown are mean ± SEM for $N = 10$ random network simulations at each point. Columns show the proportion of dead units (left), the average number of place fields per active unit (middle), and the mean peak rate of place fields (right). A. Statistics across λ holding $J_{\text{inh}} = 0.0$ (top) and $J_{\text{inh}} = 3.0$ (bottom). B. Statistics across J_{inh} holding $\lambda = 1.5$.

ing, it may not be possible to tune the output for biologically relevant spatial representation. Enabling strong recurrent inhibition by setting $J_{\text{inh}} = 3.0$, both sparsity and N.F. are quasi-linearly and predictably modulated by λ such that there are relatively broad parameter ranges where good spatial coding is available (Figure 3.12a, bottom). Interestingly, in this case, average peak rates are non-monotonic with λ (see Figure 3.15 and Section 3.4.5). Fixing λ to the value max-

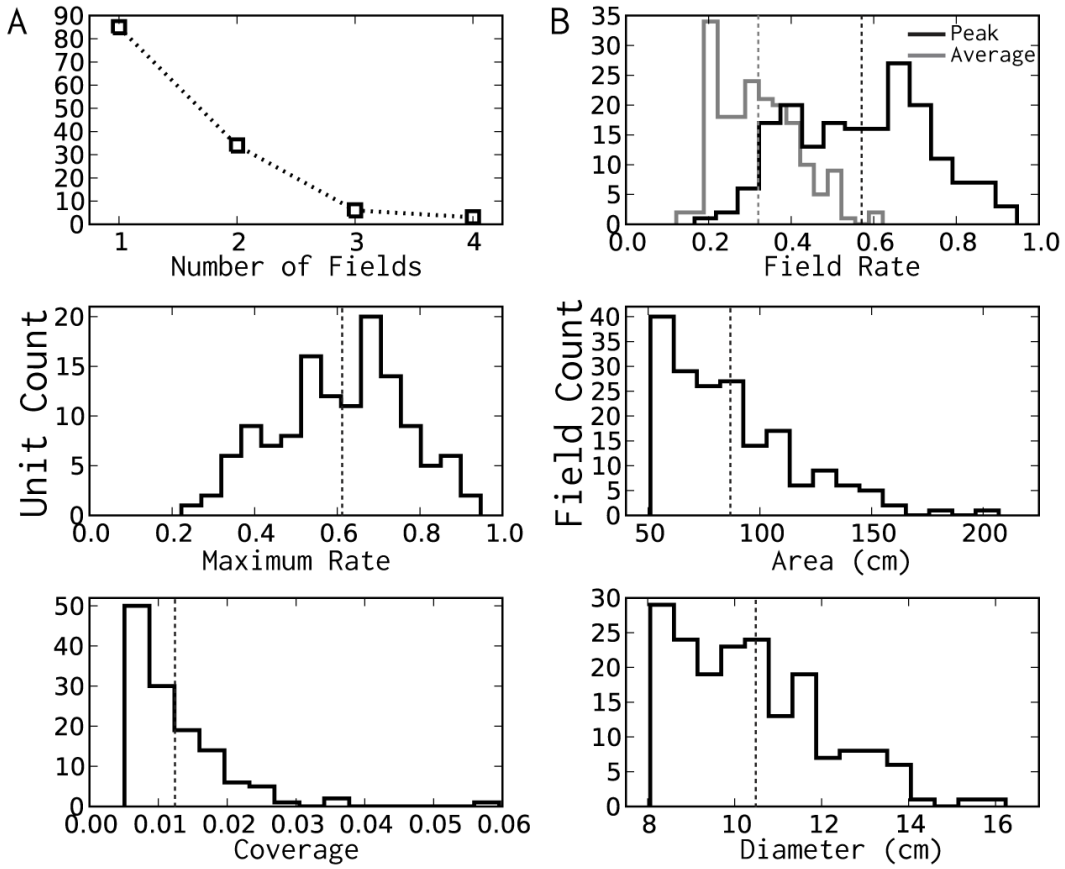

Figure 3.13 Distributions of spatial map characteristics for a single simulation at $J_{\text{inh}} = 2.5$, $\lambda = 1.5$. A. Unit-based measurements (out of 128 active units). B. Place field-based measurements (out of 183 place fields). Note that field diameter is computed directly from field area assuming circularity. Vertical dotted lines indicate means.

imizing average peak rates, $\lambda = 1.5$, the J_{inh}-dependence of these key measures of spatial coding is similarly robust and predictable, but effectively monotonic (Figure 3.12b). Here, the value $J_{\text{inh}} = 2.5$ minimizes N.F. to approximately 1.4 while also producing network sparsity of 55%.

3.4.4 Statistics of network spatial maps

We have shown that the parameter reference-point of $J_{\text{inh}} = 2.5$ and $\lambda = 1.5$ produces sparse and informative spatial maps with biologically relevant properties

and well-distributed place fields. We can then assess the individual distributions of place coding properties that generate the averages and statistics that we discussed above. The number of fields is distributed exponentially: 85, 34, 6, and 3 place units have 1, 2, 3, and 4 place fields, respectively (Figure 3.13a, top). This averages to 1.43 fields per unit. Unit maximum rates (Figure 3.13a, middle) and place-field peak rates (Figure 3.13b, top) are normally distributed across most of the possible range of firing rates, from 0.2–1.0. Finally, the spatial extent of active response is skewed toward the low end, such that larger fields are less likely to form. This is evident in the distributions of the total proportional coverage by place units (Figure 3.13a, bottom) and place field size (Figure 3.13b, bottom panels). Population means are indicated by vertical dotted lines.

The simulated spatial maps presented here depend on a random network and a random environment, so these place code distributions will vary across simulations. To assess the degree of variance in the spatial coding properties of reference-point simulations, we simulated sample sets of 25 spatial maps for several different conditions: random network paired with a random environment; random networks paired with the same environment; and a single network paired with random environments. For each spatial coding characteristic, we computed means, CIs, and SDs (Table 3.1). The SDs for the single-network and single-environment sample sets are shown alongside the data for the random sample sets. Even though both random networks and environments contribute variance to the statistics of spatial coding around the reference point, overall spread is low.

3.4.5 Dynamic competition for representation

The competitive network dynamics mediated by recurrent inhibition enable the uniform distribution of place fields across the environment despite inhomogeneities

Table 3.1 Spatial map statistics for random samples at the parameter reference-point for balanced spatial coding. 'Map' values are computed across sampled maps; 'Units' across all active place units in sampled maps; and 'Fields' across all active place fields in sampled maps. Mean, CI (95% confidence interval), and SD (sample standard deviation) describe a sample set of random environments paired with random networks. Net. SD is for a sample set of random networks against a single environment; Env. SD is for a single network against random environments. For all sample sets, $N = 25$.

Type	Value	Mean	±	95% CI	SD	Net. SD	Env. SD
Map	Sparsity	0.548	±	0.0107	0.0272	0.0311	0.0187
	Coverage	0.935	±	7.10E–03	0.0181	0.0124	9.02E–03
	Representation	1.723	±	0.0594	0.152	0.146	0.0503
	Max. Rate	0.962	±	7.40E–03	0.0189	0.0237	0.0131
Units	Num. Fields	1.438	±	0.0244	0.726	0.746	0.749
	Coverage	0.0127	±	2.63E–04	7.82E–03	8.22E–03	8.28E–03
	Max. Rate	0.621	±	5.87E–03	0.174	0.172	0.185
Fields	Area (cm)	88.4	±	0.899	32.0	32.4	32.6
	Diameter† (cm)	10.5	±	0.0509	1.81	1.82	1.84
	Peak Rate	0.583	±	4.99E–03	0.178	0.176	0.189
	Average Rate	0.327	±	3.01E–03	0.107	0.107	0.112

† Field diameter is computed directly from area assuming circularity.

in the input norms. This is evident, for the reference-point spatial map, in both its representation map (Figure 3.7b) and output norms (Figure 3.14b). Unlike the sparsity-matched threshold case, the input and output norms are significantly but weakly correlated (per-pixel Pearson $r = 0.19$, $p < 10^{-84}$; Figure 3.14c). The activity-dependent competitive mechanism enables the network to mitigate the tracking of input fluctuations and instead develop spatial codes that rely on the detailed relative distribution of the input currents at any point in the environment. This input independence is further illustrated by the marginal distributions: the strongly peaked MEC norms are transformed into output norms that are dis-

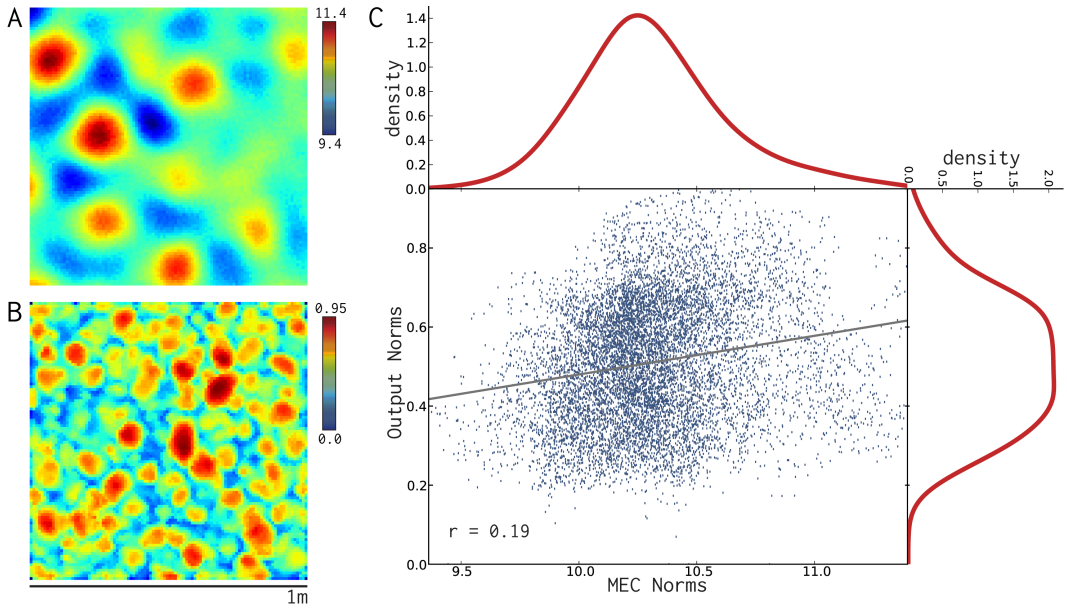

Figure 3.14 Input and output fluctuations across the environment are not strongly coupled. A,B Population rate vector norms for each pixel in the environment. A. MEC input set (same data as Figure 3.10a). B. Spatial map output. C. The MEC input norms are significantly but weakly correlated with the output norms ($r = 0.19$, $p < 10^{-84}$), for a simulation at $J_{\text{inh}} = 2.5$, $\lambda = 1.5$. Scatter plot shows per-pixel input–output comparison with regression line. Marginal distributions are smoothed with a Gaussian kernel for visualization (see Methods). The strongly peaked input distribution yields a broad, flat output distribution, indicating that the network is able to mitigate input-following by developing spatial codes based not on the magnitude of MEC inputs but on their detailed relative distributions across the environment.

tributed broadly and uniformly (Figure 3.14c).

If recurrent inhibition mediates a dynamic competition for representational power, then stronger fields should represent relatively more space. To assess this, we can correlate the peak rate of a place field with a measure of the distance of its neighbors: here, we use the mean COM distance of the two nearest neighbors. There is significant peak–neighbors correlation for the $\lambda = 1$ (slope $(m) = 1.2$ cm/rate, $r = 0.22$, $p < 0.001$) and $\lambda = 1.5$ ($m = 3.9$, $r = 0.47$, $p < 10^{-10}$) cases, but not for $\lambda = 2$ ($m = -0.47$, $r = -0.02$, N.S.). This implies an upper bound on λ

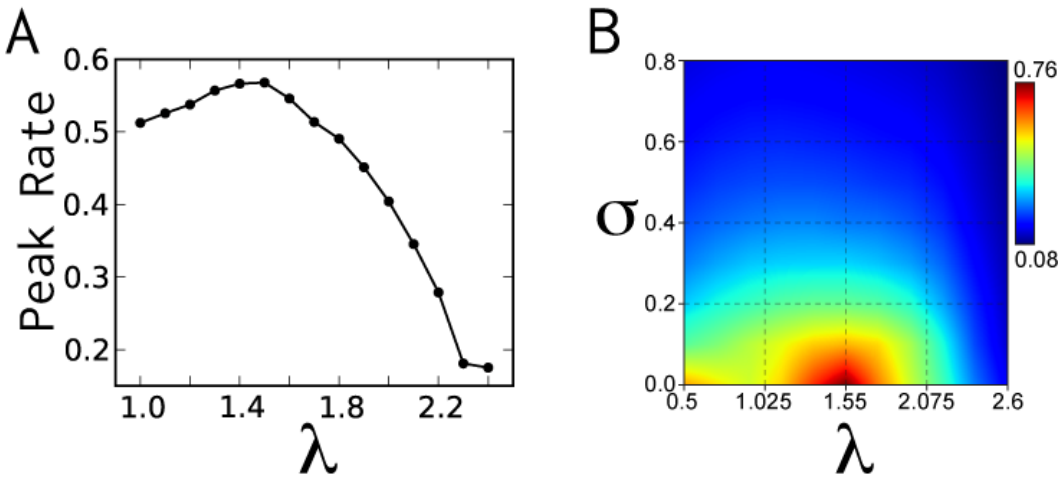

Figure 3.15 Average place-field peak rates are inverse U-shaped with λ when recurrent inhibition is strong. A. Peak rates are maximized by a moderate threshold λ value (same panel as Figure 3.12a, bottom right), but only when recurrent inhibition is active ($J_{\text{inh}} = 3.0$). Here, the smoothness of the nonlinearity is fixed at the default value of $\sigma = 0.1$. B. Parameter sweep (see Methods) showing the mean peak rate across the position (λ) and the smoothness (σ) of the place-unit nonlinearity (Equation 3.1). Peakiness of the λ curve is strongly attenuated by increasing σ.

for enabling representational competition.

Spatial map statistics generally demonstrate monotonic dependence on the dynamic parameters (Figures 3.3a, 3.4, and 3.12). As mentioned above, one exception to this is the mean place-field peak rate across λ, which is inverse U-shaped with a maximum around $\lambda = 1.5$ when $J_{\text{inh}} = 3.0$ (Figure 3.15a). The $(J_{\text{inh}}, \lambda)$ parameter sweeps further show that this peak occurs in the region around the reference point (Figures 3.3b and 3.4c). If this non-monotonicity is indicative of cooperative or balanced network activity, then it is instructive to assess how it may be modulated by the other parameter of the place-unit nonlinearity (Equation 3.1), its smoothness σ, which has been fixed so far to the value $\sigma = 0.1$. We computed a (λ, σ) parameter sweep, holding inhibitory gain fixed to $J_{\text{inh}} = 3.0$, demonstrating that the λ-dependence of the mean peak rate is strongly modulated by σ (Figure 3.15b).

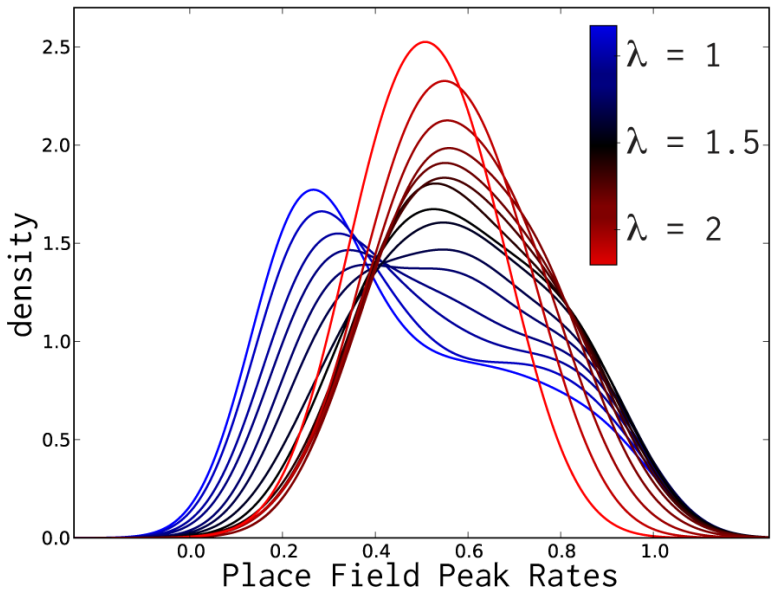

Figure 3.16 Modulating the amount of afferent excitation affects the quality of the peak rate distribution. Increasing λ broadens the peak rate distribution up to $\lambda \simeq 1.5$, after which the distribution becomes more Gaussian and the mean peak rate drops. Densities are smoothed with a Gaussian kernel for visualization (see Methods).

Decreasing σ below $\sigma = 0.1$ increases both the peakiness of the inverse U-shape and the maximum value of the mean peak rate. This indicates that a more step-like nonlinearity may further enhance the competitive balance; in effect, enabling cortical excitation to more efficiently drive the network. However, we observed that using a Heaviside step function as the place-unit nonlinearity results in non-Gaussian place fields and rate distributions that are not capable of maintaining an informative rate code (data not shown). Much larger values of σ severely diminish λ-modulation of the mean peak rate. Lastly, because this maximum coincides with the high peak–neighbors regression slope and correlation at $\lambda = 1.5$, the λ effect on representational competition may also be evident in the λ-dependence of the population rate distribution.

To assess the comodulation of competition and peak rates, we simulated an-

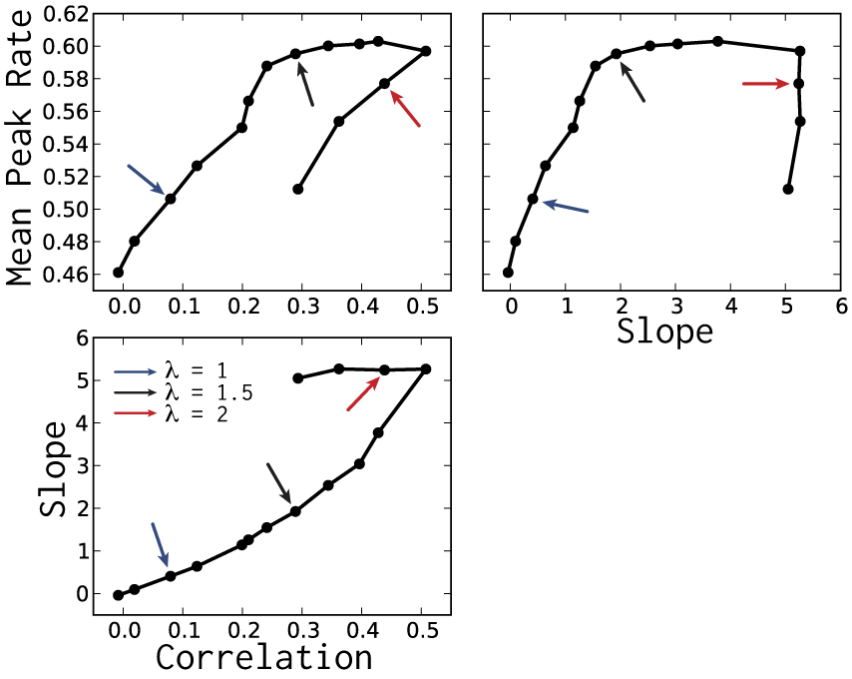

Figure 3.17 Moderate threshold enables both effective representational competition and broadly distributed peak rates. Peak–neighbors regression slope and correlation vary smoothly together and with mean peak rate as threshold increases from $\lambda = 0.8$ to $\lambda = 1.8$, but at higher activity thresholds there is a disjunction in the quality of competitive spatial coding. The reference-point $\lambda = 1.5$ simulation (black arrows) maximizes mean peak rate and has stronger competitive slope and correlation than lower thresholds. This suggests a balance between J_{inh} and λ that enables both healthy competition and coherent spatial representation.

other network–environment pair holding J_{inh} fixed at 2.5 while varying λ from 0.8 to 2.2. The peak rate distribution of the whole place population changes significantly across this range (Figure 3.16). For low thresholds (blue lines), there is a strong mode at the low end with a long tail to the high end of the range, indicating that a small subset of winners is suppressing the rest of the population. For moderate thresholds around $\lambda = 1.5$ (black lines), corresponding to the maximum of the mean peak rate across λ (Figure 3.15), the distribution has broadened and become more symmetric, indicating a more uniform rate distribution. Further increasing λ beyond 2.0 (red lines), the rate distribution narrows substantially and shifts lower.

That is, if λ is too low, a small number of units win everything; too high and the population is relatively equal but poorly tuned. It also changes the quality of competition: the peak–neighbors regression slope and correlation vary together and with the mean peak rate smoothly up to $\lambda \simeq 1.9$, where the progression of representational competition and the rate distribution changes (Figure 3.17). We suggest that these effects of λ, by modulating the amount of input activity driving the competition, demonstrate that a balance between the afferent excitation and the recurrent inhibition enables competitive fairness across the widest dynamic range. Thus, our parameter reference-point of $J_{\text{inh}} = 2.5$ and $\lambda = 1.5$, which we previously demonstrated to produce qualitatively relevant sparsity and spatial specificity (Figure 3.13), also enables efficient, effective and balanced competition for representation of the environment.

3.4.6 Autocorrelations and spurious place fields

Secondary place fields, those active fields of a place unit with peak rates less than its maximum rate, contribute to decreased spatial specificity for that unit and spatial coherence for the map as a whole. Inspection of the population autocorrelations of the simulated MEC inputs and place-unit response outputs reveals where these fields tend to appear. The periodic structure of the grids is evident as a ring of secondary correlations around the central peak in the MEC population autocorrelogram (Figure 3.18a, top). The peak of this ring occurs at a radius of around 40 cm (Figure 3.18a, bottom), closely corresponding to the median grid spacing in our simulations of 39.2 cm. Notably, a qualitatively similar correlation structure is evident in the autocorrelogram of the output spatial maps, though at a much reduced level relative to the central peak (Figure 3.18b). Since input peaks across place units are randomly distributed in the environment (due to sparse, randomly weighted

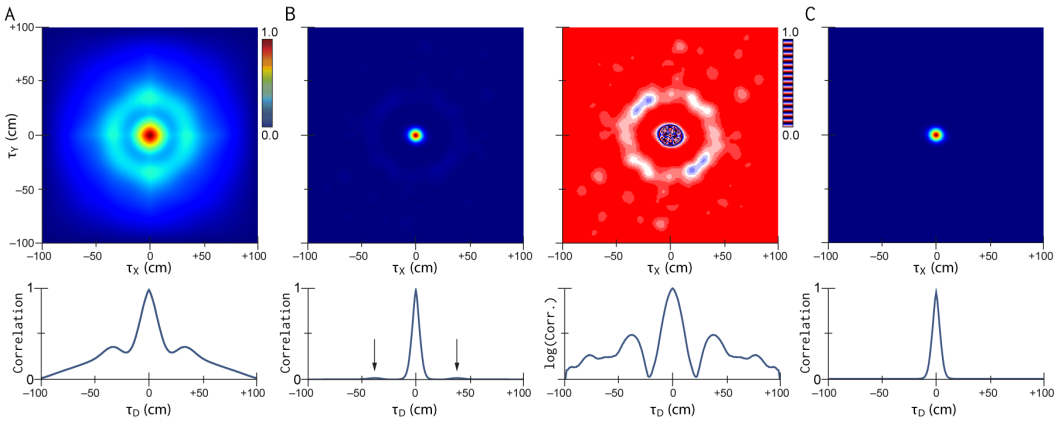

Figure 3.18 Autocorrelations for the simulated MEC inputs and spatial map outputs. (top) Two-dimensional population autocorrelograms. (bottom) Linearizations of the corresponding autocorrelograms. A. The simulated MEC input set shows strong secondary correlations off the central peak at the median grid spacing 39.2 cm. **B,C** Spatial map outputs. The width of the central peak at half-max is 7.9 cm. B. The grid-based secondary correlations feed through, though at a low level (arrows, left). Viewing the same data with a periodic colormap and log plot (right) enhances the contrast of these residual correlations for visualization. C. Restricting the population ratemap to just the primary place fields removes the grid-dependent correlations, indicating that secondary place fields tend to develop at the grid vertices of strong inputs that also contribute to the primary field.

synapses from an input set with random spatial phases), it is not likely that these input-driven secondary correlations are due to correlations between place units. There must be some degree of grid-scale correlation within the response maps of individual place units. To assess this, we considered the distribution of the COM distances of secondary place fields from their unit's primary field. This distribution, computed for the 1,485 total secondary fields in the 25 spatial maps of the reference-point random sample (Table 3.1) and smoothed with a Gaussian kernel (see Methods), has a strong proximal mode at 39.0 cm and a weaker second mode at 78.0 cm from the primary place field (Figure 3.19). The shallowness and high dispersion of the second mode may be due to frequency-mixing, the diversity of ver-

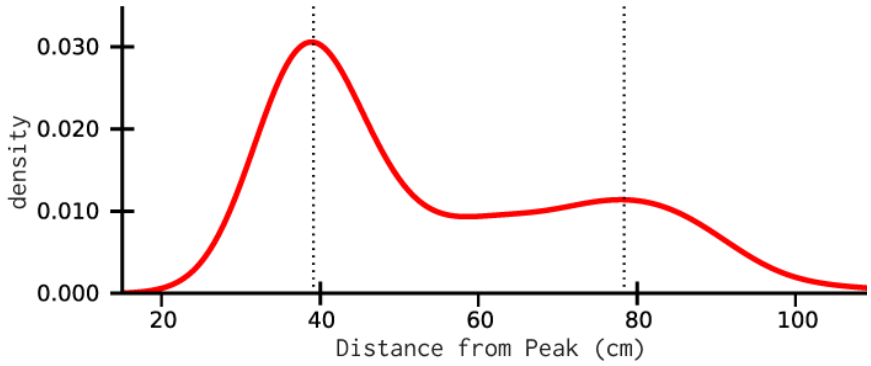

Figure 3.19 Spurious fields tend to form at the vertices of strong grid inputs. The distribution, for all 1,485 secondary (non-primary) place fields in the sample of 25 random reference-point simulations (Table 3.1), of COM distances from each place unit's respective primary field. Modes occur at both 39.0 and 78.0 cm from the primary field COM, corresponding to multiples of the median grid spacing of our set of simulated MEC inputs (39.2 cm). Since there are no grid-scale correlations in spatial phases between grid maps, this indicates that spurious fields tend to form at vertices of strong grid inputs that also contribute to the cortical input peak underlying the primary place field. The distribution is smoothed with a Gaussian kernel for both visualization and mode determination (see Methods).

tex distances in the hexagonal MEC grid structure, and/or frequency-dependent decorrelation of grid spatial phases with distance from the primary place field. However, these modes occur at nearly exact multiples of the median grid spacing, indicating that secondary fields tend to form at the vertices of grid inputs that contribute to the input peak that gives rise to the place unit's primary place field. Indeed, if we construct output ratemaps that set to zero all values outside of a unit's primary field, then the grid-scale secondary correlations disappear leaving only the central peak of the autocorrelation (Figure 3.18c). This peak has a width of 7.9 cm at 50% and 13.4 cm at 20% of the maximum correlation.

3.4.7 Onset dynamics of map formation

The competition among place units to represent the environment occurs through time, so we can examine the temporal evolution of spatial coding as the network initially responds to a novel environment. In our simulations, the amount of time equal to a single integration time-constant τ is divided into 10 discrete computational time-steps and each pixel is clamped for typically 5–6 τ (see Methods). To address the time-course of response, we created a sequence of spatial maps for a random network–environment pair corresponding to each simulation time-step (see Methods) from 0.0 to 5.3 τ. So that each pixel in the environment initializes its response in the same state, the population rate vector was reset to zero as the network was clamped to each new input vector. Due to the fixed nonlinearity and recurrent dynamics of the model, network rates only become significantly nonzero after a certain period of time, typically 2–2.5 τ. For this series of spatial maps, this occured at 2.4 τ, so we consider the dynamics of the response from that point on using the index Δt, ranging from 0.0 to 2.9 τ.

Several measures of spatial representation converge rapidly. Initially, all place-unit responses are active across the environment, but rates are below the noise floor. Thus, sparsity decreases from 1.0 and N.F. increases from 1.0 as the interneuron ramps up during the response, driving the sparsification of the outputs. Those values, along with environment coverage and per-pixel representation, rapidly converge within 1–1.5 τ of the onset of suprathreshold activity (Figure 3.20a). The absolute maximum rates of the place units are relatively slow to develop during the response (Figure 3.20b, dotted lines), but the structure of their distribution (normalized to the spatial map maximum) is largely in place by $\Delta t = 2.0\,\tau$ (Figure 3.20b, solid lines). The rate maxima do not follow the same temporal evolution across the population. We can sort the place units by their final maximum

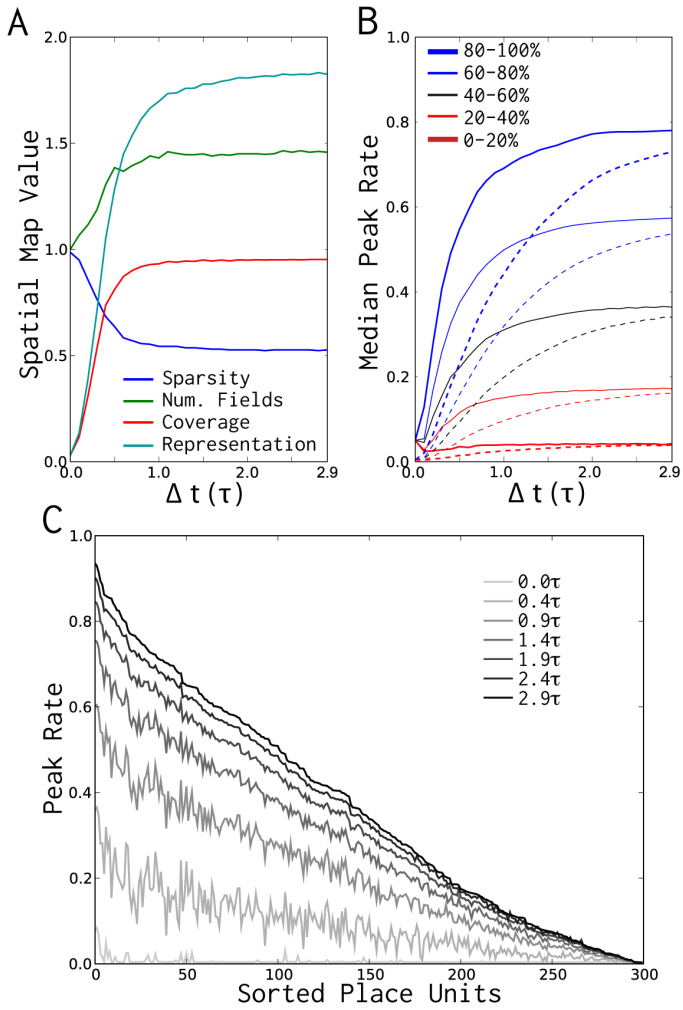

Figure 3.20 Short-timescale dynamics of spatial map formation. This data is based on a series of spatial maps created at each time-step as a random network is clamped to a novel environment (see Methods). This progression is indexed by Δt which varies from 0.0 to $2.9\,\tau$, indicating time (expressed in integration time-constants) since the initial onset of suprathreshold activity. For our simulations, there are 10 discrete time-steps per simulated τ. A. Time-course of four critical measures of spatial coding shows stabilization around $1.0\,\tau$. **B,C** Peak rates sorted by final peak rate at $\Delta t = 2.9\,\tau$. B. Evolution of median peak rate of place-unit quintiles. Solid lines are normalized to map maximum rate; dotted lines are absolute rates. C. Peak rates for all $N = 300$ place units. The sort does not hold for earlier points in the response, indicating that the heterogeneity of competition across the environment differentially influences the time-course of the rate response.

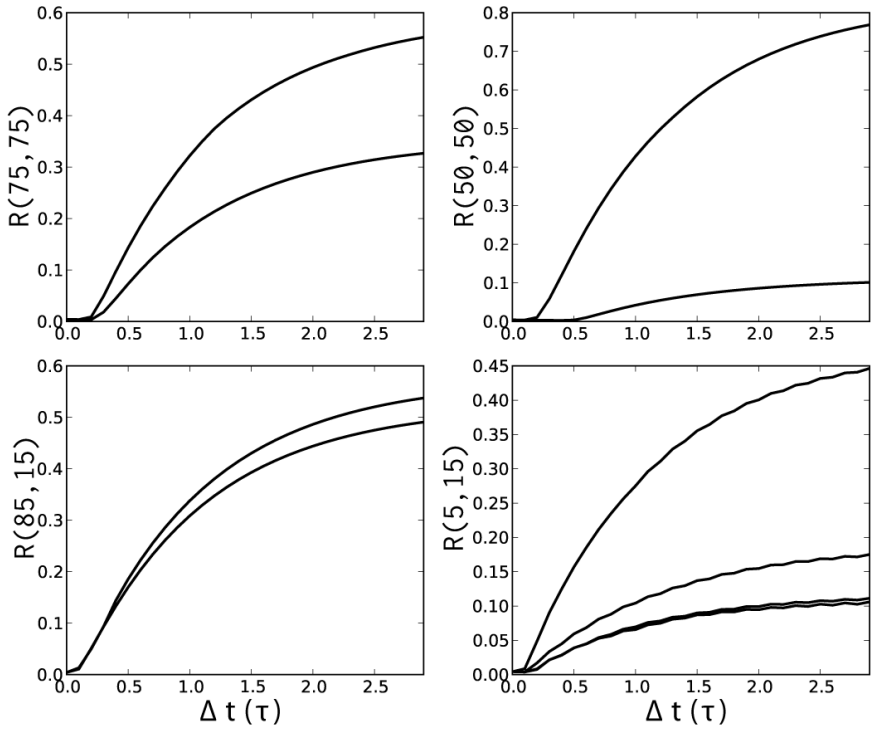

Figure 3.21 Examples of rate evolution of winning units at single locations in the environment. The onset of winning place-unit rates are shown for four different locations in the environment as the network response progresses. Only units with final rates greater than 0.1 are shown. The detailed distribution of afferent input currents across the population determines how well a given place unit can compete at a given location. This heterogeneous competitive balance contributes to differential rate evolution (Figure 3.20c) and to dynamic pattern separation by enhancing input sensitivity.

rate at $\Delta t = 2.9\,\tau$ and show that the sort does not hold for progressively earlier periods in the response (Figure 3.20c). Though, there remains a tendency for units with relatively high maximum rates early in the response to also converge to relatively high rates. The rate evolution is dictated by the detailed competitive balance at the particular location in the environment to which a given place unit ends up responding maximally. The rate response for the winning units at example locations (Figure 3.21) shows that the competition usually occurs between a small number of units and the particular rate evolution of those units is unique to the

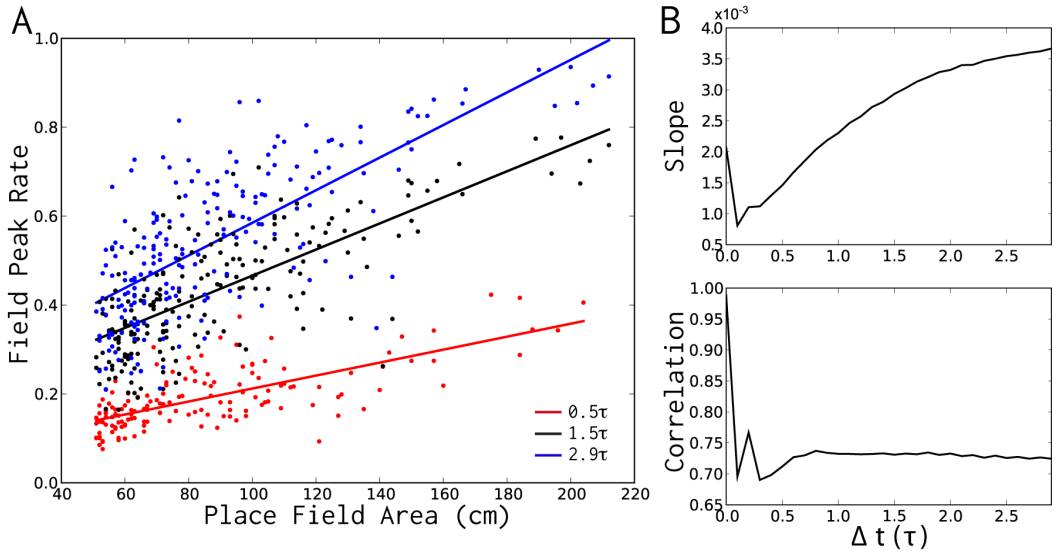

Figure 3.22 The place field area–peak relationship strengthens with the response. A. Per-field scatter with regression lines at several points of the response. B. The area–peak slope increases steadily (top) and maintains a strong correlation around $r = 0.72$ (bottom).

location. Every location has its own detailed distribution of afferent input currents (Equation 3.4) that determines the time-course of the place competition.

If place fields are considered Gaussian firing fields of a certain spatial scale, then there should be a positive correlation between the peak rate and size of a place field. Indeed, significant area–peak correlations are evident throughout the response (Figure 3.22a). The slope of this relationship steadily increases with the response while maintaining high correlation (Figure 3.22b). The rate of increase of the slope levels off by $\Delta t = 2.9\,\tau$ but continues to increase. The range of place field areas broadens somewhat during the response, so the increase in slope is due primarily to differential evolution of the peak rates: stronger place fields increase their peak rates faster.

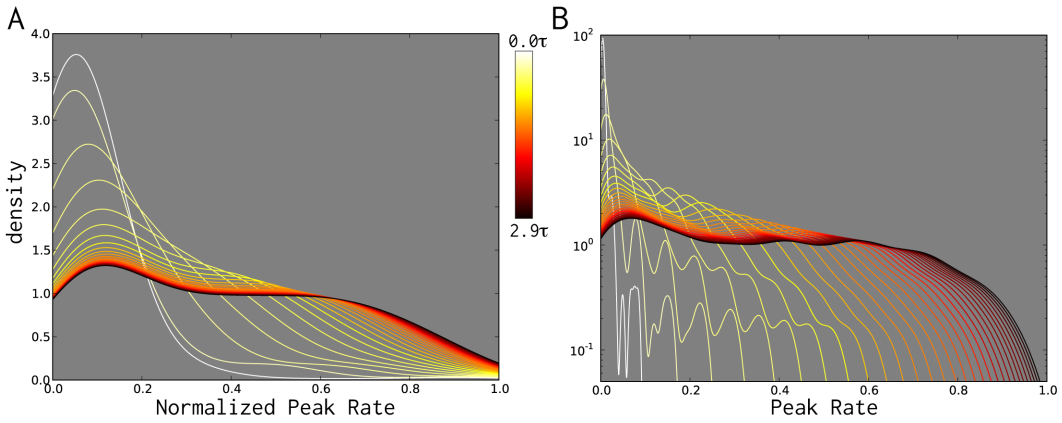

Figure 3.23 Evolution of peak rate distributions shows rapid convergence in the relative structure of the rate code. Population rate distributions initially only have a very small number of units with non-zero rates. This low mode rapidly broadens out and converges to become more uniformly distributed as the response progresses. A. Rates normalized to maximum map rate. B. Absolute peak rates (shown as log plot). The distributions are smoothed with a Gaussian kernel for visualization (see Methods).

3.4.8 Dynamics of place and rate code convergence

Due to its importance to the resultant spatial coding characteristics, we can examine the temporal evolution of the population rate distribution. It can be indicative of both the timescale of rate-code convergence and the effects of competition on representational strength over time. The relative rate distribution (rates normalized to the spatial map maximum) changes rapidly over the course of the response (Figure 3.23a). Initially, at $\Delta t = 0.0\,\tau$, most units have not begun to respond and only a small subset has non-zero rates. At that point, absolute rates only range from 0.0–0.1 (Figure 3.23b). This initial low mode begins to broaden into a more uniform distribution within a small number of time-steps. Around $\Delta t = 1.0\,\tau$, or 10 simulation time-steps after the onset suprathreshold activity, the relative rate distribution does not appear significantly different from the final distribution at $\Delta t = 2.9\,\tau$ even though the range of absolute rates is only 75% of

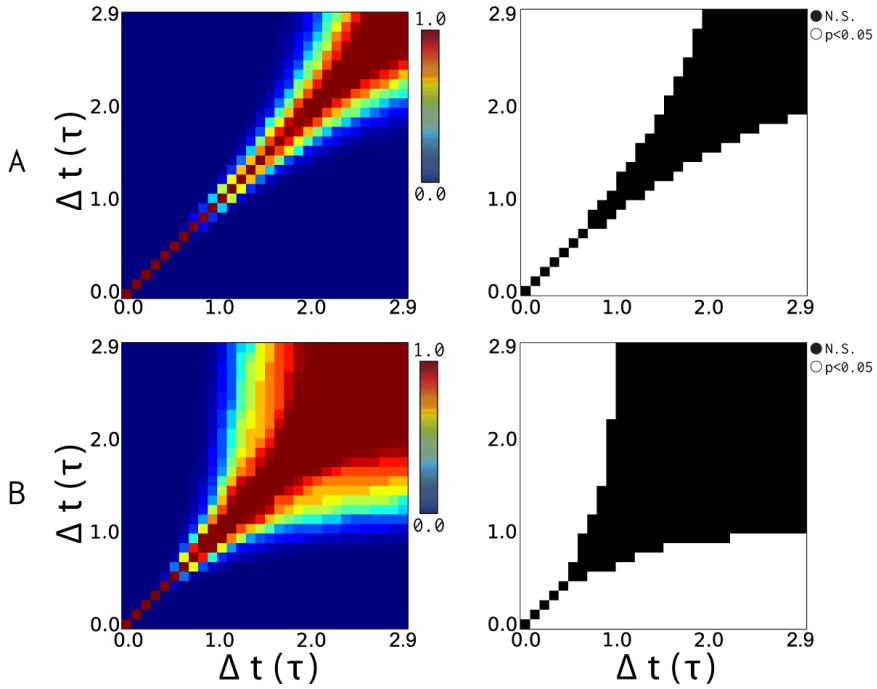

Figure 3.24 Population rate distributions become statistically different on the timescale of integration. Pair-wise independent Kolmogorov-Smirnov (K-S) tests for changes in peak rate distributions across the response. K-S test p-values (left) and significance at $p = 0.05$ (right; black indicates N.S.) show that the absolute rate distributions (A) converge slower than the normalized rate distributions (B), similar to but overall slower than the network average changes shown in Figure 3.25. The absolute rates lose significance at $\Delta t = 1.9\,\tau$ while the normalized rates do so at $\Delta t = 1.0\,\tau$.

its final range (Figure 3.23). This indicates that the relative structure of the rate distribution converges on a faster timescale than the absolute rates. This is verified by pair-wise Kolmogorov-Smirnov tests between each time-step of the response. The absolute rates first become statistically similar to the final distribution ($p > 0.05$) at $\Delta t = 1.9\,\tau$ (Figure 3.24a) while the normalized rates do so at $\Delta t = 1.0\,\tau$ (Figure 3.24b). The final distribution is fairly broad: there is a small local peak below 20% of maximum, a flat plateau from 20–60%, and then the density approaches zero as rates approach the map maximum (Figure 3.23). As the network responds to its cortical afferents, the rates tend to increase on the timescale

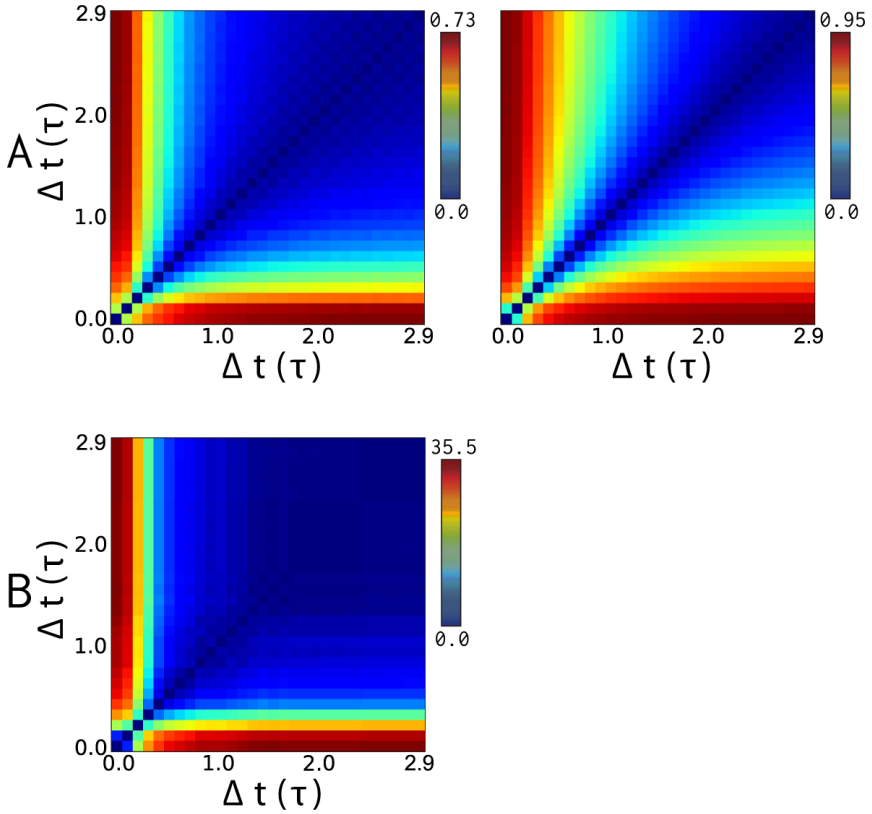

Figure 3.25 Pair-wise convergence of place-unit peak rate and location during the response. Population changes in place and rate coding are quantified between every pair of time-steps in the response. A. Rate codes: network average of peak-rate change for max-normalized (left) and absolute (right) rates. B. Positional codes: network average of peak firing location distances across place units. The positional codes converge almost immediately, followed by the normalized rate structure and then the absolute rates.

of their integration constant τ, but this growth in place-unit activity equivalently drives the global interneuron. This balanced inhibitory drive increases both sparsity and spatial specificity as the network response progresses.

Does this process differentially affect the timescale of availability of place and rate codes? To assess this, we can consider pair-wise measures of changes to place unit maximal rates and locations for each time-step across the response. For the rate code, the network average of the normalized rate change for normalized rates

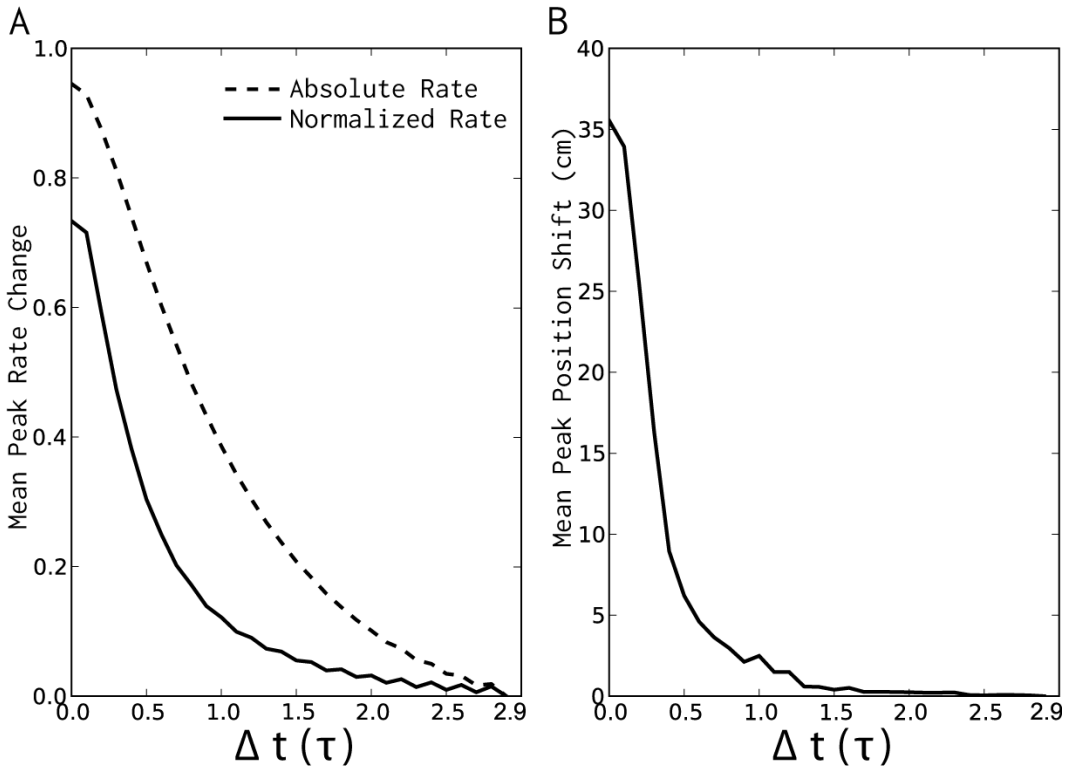

Figure 3.26 Differential time-courses for population covergence of place and rate codes. Population changes in spatial coding (as in Figure 3.25) are shown relative to just the final peak rates and locations. A. Half-max convergence of the normalized (solid line) and the absolute (dotted line) rate code occurs in $\Delta t = 0.42\,\tau$ and $\Delta t = 0.82\,\tau$, respectively. B. Half-max convergence of the positional code occurs in $\Delta t = 0.28\,\tau$. The place code converges much faster than the rate code, and it is the first component of the spatial representation to become available.

converges to zero faster than that of absolute rates (Figure 3.25a). Considering the rate change from just the final time-step, $\Delta t = 2.9\,\tau$, half of the maximum evolution has occured for normalized rates by $\Delta t = 0.42\,\tau$ (using linear interpolation between time-step values) and for absolute rates by $\Delta t = 0.82\,\tau$ (Figure 3.26a). A measure of positional change, the population average of distances that maximal response locations shift, shows pair-wise convergence within a very small number of time-steps (Figure 3.25b). Indeed, half-max evolution toward the final positional code has occured by $\Delta t = 0.28\,\tau$ (Figure 3.26b), which is 33% faster than absolute

and 66% faster than normalized rate-code convergence, respectively. Thus, the structure of the final spatial code is largely available within 1–1.5 τ: the positional code is nearly immediately available while the relative and absolute rate codes converge relatively more slowly.

3.4.9 Trajectory dependence of spatial coding

Having analyzed the properties of spatial coding for this model, it is important to determine whether the network response ratemaps that we create based on raster scans of the environment (Section 3.3.3) are representative of the output of behavioral trajectories. To address this, we simulated the spatial map of a random network–environment pair as usual and compared it to the real-time output of a naturalistic trajectory. This 30-s trajectory is a smooth random walk with an average linear speed of 15 cm/s (Figure 3.27a) to emulate the exploratory or foraging behavior of a rat in a novel environment. For such a trajectory, the value of the integration time-constant (Equation 3.1) influences the output since it determines the relative responsiveness of the network to the continuously changing cortical inputs. Relatively fast integration should result in less discrepancy between raster-based and real-time output by more closely tracking input changes and allowing the place-unit competiton to converge faster (Figure 3.26).

Because the time-scale of a putative biological analog for the inhibitory feedback interaction that we model here is at best unclear (Section 3.5.3), we simulate the naturalistic trajectory for five values of τ, ranging from 50–250 ms. All of these simulations were integrated in 5 ms time-steps. At every time-step, we show the correlation strength between the population rate vectors of the output from the naturalistic trajectory and the corresponding location in the pre-computed ratemap (Figure 3.27b). Correspondence between the two spatial codes is generally high

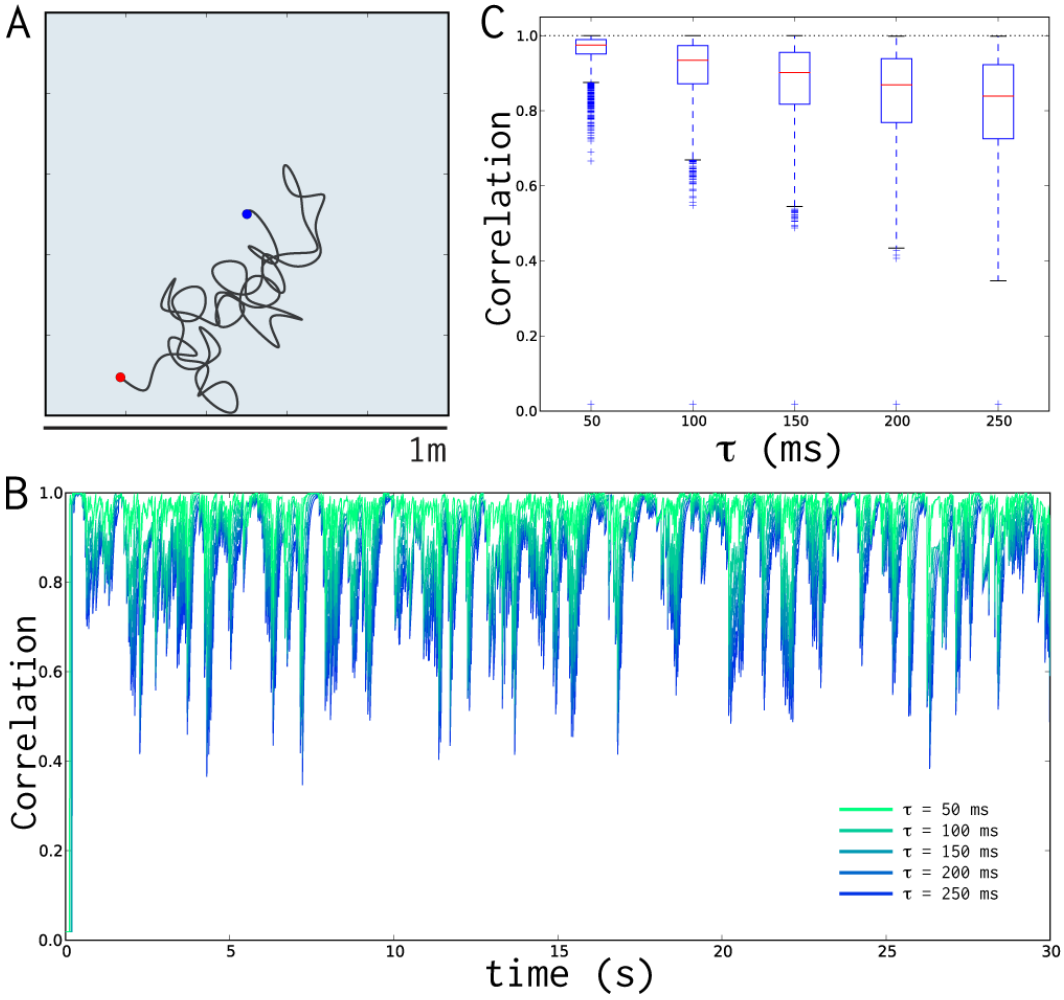

Figure 3.27 Spatial map properties determined using an artificial raster trajectory are strongly correlated with outputs from a naturalistic trajectory. We simulated the spatial map for a single network–environment pair as usual (Section 3.3.3) and compared it to the real-time output of a smooth random-walk trajectory (average speed 15 cm/s). A. The 30-s trajectory through the 1m square environment (blue circle: start location; red circle: end location). B. Correlation strength between corresponding population rate vectors in the raster-scan spatial map and the random-walk output across time. The naturalistic trajectory was simulated with five different values for the integration time-constant τ (Equation 3.1), ranging from 50–250 ms. C. Box-and-whiskers plots of the distribution of trajectory correlations across time for each τ.

but characterized by irregular, transient decorrelations. Correlations between these events approach $r = 1$. The time-course is similar across τ, with slower integration

resulting in more substantial decorrelations as expected. Distributions of correlation across time (Figure 3.27c) show that, even with longer τ, there is high correlation during most of the random-walk trajectory. Median correlations (red lines in Figure 3.27c) range from $r = 0.84$ for $\tau = 250$ ms to $r = 0.98$ for $\tau = 50$ ms. This makes sense considering that the spatial scale of network output is about 13 cm (Figure 3.18c), commensurate with the distance traversed in this trajectory in 1-s. That is, τ approaching 1-s would likely result in output unrelated to the converged responses, but then our putative feedback interaction would be outside the range of activity-driven network dynamics.

3.5 Discussion

3.5.1 A model of dynamic competition

We present a network model of hippocampal spatial representation that is both capable of global remapping (Chapter 4) and is dependent only on short-timescale network dynamics. This minimal dynamic model produces sparse, informative, population spatial maps for any particular spatial phase and orientation alignment of a simulated population of MEC grid-cell responses (Hafting et al., 2005). Here, any such alignment determines the cortical representation of an environment, all of which are novel to the network due to random afferent synaptic weights. Our interest in such a model rested primarily on the observations that MEC realignment and global remapping in CA3 are both contiguous and nearly instantaneous (Fyhn et al., 2007). Both observations put strict constraints on putative mechanisms linking the two phenomena and, further, suggest that hippocampus is performing a dynamic real-time readout of its cortical inputs.

The hypothesis of a real-time readout is consistent with a number of recent studies showing that the perforant path structures receiving MEC layer II input, dentate gyrus (DG) and CA3, form an independent spatial coding system specialized for indexing distinct environmental contexts (Vazdarjanova and Guzowski, 2004; Leutgeb et al., 2004; Lee et al., 2004; Lee and Knierim, 2007) and rapidly integrating new spatial information (Lee et al., 2004; Leutgeb et al., 2006; Leutgeb and Leutgeb, 2007). Much of the rapid spatial learning in DG and CA3 depends on synaptic plasticity (Nakazawa et al., 2003; McHugh et al., 2007; Leutgeb et al., 2007; Haberman et al., 2008), but to achieve nearly instantaneous readout of the structure of MEC responses, we restricted the model to activity-dependent dynamics. That is, the spatial codes developed by this model comprise the initial state of the network upon introduction to a novel environment; the resulting "prototype maps" may drive subsequent processes of synaptic modification that evolve the spatial maps with further experience in the environment. Because the temporal and metabolic costs of learning are related to the degree to which a prototype map must be modified in order to become functional to the animal, our goal was to achieve qualitatively relevant levels of sparsity and spatial specificity in the individual responses and robust representation of the environment in the population maps. That is, better prototype codes enable faster, more efficient evolution of the familiarized representations attained through learning.

To approach this question, then, we employed nonlinear place units and feedback inhibition provided by a single global interneuron. To ensure that any heterogeneity in output responses are due to the network interaction, the place population is homogeneous up to the particular permutation of afferent synaptic weights from the simulated MEC grid-cell response maps (see Methods). Thus, we allow the feedback inhibition to mediate dynamic, activity-dependent competition

among the place units as the sole source of response diversity. For the cortical inputs, we made no assumptions regarding metric structure of the grids in MEC, so each simulated grid map has a random spatial phase and orientation with grid periodicity sampled from the observed range (see Methods).

The spatial frequency range of MEC inputs used here is associated with CA3 spatial responses having the lowest relative spatial scale of correlations on a large track (Kjelstrup et al., 2008), so network output may correspond to the scale evident along the most dorsal 20% of hippocampus. Since hippocampal pyramidal cells typically are innervated by projection cells along a dorsoventral extent of at least 25% of MEC (Dolorfo and Amaral, 1998) and the relative scale of grid-cell responses tends to cluster into subgroups (Barry et al., 2007), we suggest that it is likely that this frequency range corresponds to a single projection band to hippocampal areas near the septal pole. As such, this input–network pair abstracts a dorsal hippocampal subnetwork served by a single interneuron (or a correlated set of interneurons).

3.5.2 Place representations and spatial phases of grids

Initially, we demonstrated that this mechanism can be broadly tuned, where strong inhibitory gain is coupled with balanced afferent excitation, to produce sparse and informative spatial maps. Typically, in hippocampus, around 70% of principal cells in CA3 and 50% in CA1 are silent "dead units" within a given environment (Wilson and McNaughton, 1993; Lee et al., 2004; Leutgeb et al., 2004). One study of simultaneously recorded CA3 pyramidal and DG granule cells demonstrated, on average, 1.1 and 1.9 place fields per active cell, respectively (Leutgeb et al., 2007). Here, on average, our maps exhibit 54.8% sparsity and cover 93.5% of the environment such that 1.72 place fields overlap any given location, and active units

have 1.44 place fields (Table 1). That is, we demonstrate network sparsity similar to CA1 but not as high as CA3, and spatial specificity of individual responses better than DG but worse than CA3. The prototype maps, then, could be considered degenerate CA3 maps, characterized by the presence of spurious secondary place fields (Figure 3.19) and disproportionately large active coding subsets.

Since the discovery of grid cells, several modeling studies have addressed the derivation of place fields and maps from grid cell responses. It was noted early on that a thresholded linear summation of grid maps with matching spatial phases can produce a single restricted field at the location where the grid phases coincide (O'Keefe and Burgess, 2005; McNaughton et al., 2006). Solstad et al. (2006) showed, by thresholding linearly summed grid maps after subtracting a constant inhibitory term, that maintaining identical grid phases allows a small number (10–50) of MEC inputs to produce single well-defined place fields. More recently, Hayman and Jeffery (2008) extended a similar model to include context-dependent switching of correlated inputs sets afferent to distinct dendritic branches of linear DG units. In that model, remapping is achieved by differential activation of subsets of phase-matched MEC inputs; these correlated MEC subsets drive a layer of DG units, highly correlated subsets of which exclusively innervate CA3 units. While the CA3 units show place fields and demonstrate remapping between contexts, significant synaptic reorganization of both spatial and contextual inputs is required to facilitate the necessary input clustering and connectivity patterns (Hayman and Jeffery, 2008). Such contextual gating may contribute to map retrieval in familiar environments, but it is unlikely to be relevant on the timescale of a first pass through a novel environment.

Models integrating over grids with heterogeneous spatial phases have so far relied on some form of associative Hebbian plasticity to produce sparse spatial

codes. Rolls et al. (2006) examined a competitive learning network model of DG in which recurrent inhibition was implemented as a global activity threshold chosen to produce a specified degree of output sparsity. In that model, similar to our sparsity-matching case (Figure 3.7) in that initial sparsity is achieved purely with a global activity threshold, the linear DG units demonstrated poor spatial specificity (averaging 6.81 fields) until the afferent synapses were associatively tuned with a Hebbian learning rule (averaging 2.63 fields) (Rolls et al., 2006). More recent spiking models of CA1 (Savelli and Knierim, 2007) and DG (Molter and Yamaguchi, 2008) that integrate MEC inputs with diverse spatial phases show that such learning rules potentiate a subset of MEC inputs with coincident spatial phases, leading to well-tuned place fields and high spatial specificity. These competitive learning models rely on heterosynaptic long-term depression, whether simulated (Savelli and Knierim, 2007) or by explicit synaptic renormalization (Rolls et al., 2006; Molter and Yamaguchi, 2008), to maintain effective competition for place representation.

The present work attempts to decouple output competition from learning mechanisms that, while fast, require nontrivial experience within a new spatial context. Competitive learning can decorrelate sparse place outputs (Franzius et al., 2006) by evenly distributing place fields, but we show that this can also be performed by a tightly-coupled interneuron providing feedback to nonlinear place units. Interestingly, Savelli and Knierim (2007) combined their Hebbian spiking model with a subnetwork of interneurons and also demonstrated more uniform dispersal of place activity, though it was not critical to place field formation overall. Here, we show that while an activity threshold (as in their integrate-and-fire neurons) can indeed lead to the formation of place fields, balanced inhibitory feedback provides field formation in a way that also results in an informative population spatial rep-

resentation without initially requiring learning.

Further, while assuming diversity of spatial phase and orientation, we did not introduce variance among the subfields of individual grid-cell response maps (e.g., spatial jitter and anisotropy (Franzius et al., 2006) or peak-rate variance (Rolls et al., 2006)). We wanted the heterogeneity of place coding to be a function solely of the afferent weight permutation and the network dynamics. Though, if there is intrinsic variability between grid fields, then it may facilitate the aperiodicity of place cell responses and, perhaps, the efficacy of remapping as driven by MEC realignment. Additionally, we supposed that for a "naive" network in a novel environment, there is no reason to assume a restricted number of or particular functional organization for the cortical afferents. Thus, our place units integrate several hundred MEC grid inputs (Amaral et al., 1990) with homogeneous subfields and heterogeneous spatial metrics. This is a worst-case scenario for producing place-like outputs due to nearly uniform cortical activation across the environment. Even so, we demonstrate that balanced competitive network dynamics can transform this low-information rate code rapidly into operational spatial representations.

3.5.3 Competitive network dynamics can prototype spatial codes

Evidence in mice shows that both NMDA and protein synthesis are necessary for the long-term stability, but not the short-term formation, of hippocampal spatial maps (Kentros et al., 1998; Agnihotri et al., 2004). In rat hippocampus, place fields can be observed on the first pass through a novel environment (Hill, 1978); though the response is initially unstable, it evolves and stabilizes with experience (Frank et al., 2004). Together, these results suggest that hard-wired network dynamics may have a role in forming the initial active state of hippocampal networks in a novel environment. We use the model presented here to explore the role for

feedback inhibition as a putative component of these network dynamics.

CA3 interneurons, including those located in stratum lucidum (s.l.), are driven both as feedback and feedforward units and exhibit afferent-specific synaptic transmission pathways (Acsady et al., 1998; Pelletier and Lacaille, 2008). This supports the idea that these interneurons participate as a node in two separable inhibitory circuits: that of local recurrent synaptic drive and that of extrinsic feedforward synaptic drive. Given this, there are two different possible modes of interpretation for the model presented here. First, we can suppose it represents a functional coupling of DG and CA3 spatial coding, in which the nonlinearity threshold λ abstracts feedforward mossy-fiber inhibition and the feedback inhibition modulated by J_{inh} abstracts the recurrent trilaminar interneurons of DG and s.l. interneurons of CA3. This interpretation has several caveats in that it does not deal effectively with the intrinsic pattern separation capabilities of DG (Acsady and Kali, 2007), the role of NMDA in DG for spatial selectivity in novel environments (Croll et al., 1992; Leutgeb et al., 2007), the diversity and typology of DG interneurons (Sik et al., 1997) and the strength and complexity of excitatory–inhibitory interactions between DG and CA3 (Buckmaster and Schwartzkroin, 1995; Jaffe and Gutierrez, 2007). These caveats, however, serve to illustrate the substantial scope of inhibitory mechanisms within and between these perforant path target structures that shape and control the relatively sparse output of their principal cell populations.

Second, we can suppose that direct entorhinal–CA3 collaterals of the perforant path (Witter, 2007) are afferent to a separable microcircuit consisting of CA3 pyramidals and feedback inhibition mediated by recurrent drive of s.l. or other interneurons. In this case, λ would represent either a constant local inhibition or the intrinsic membrane excitation threshold of the pyramidal cells. There is evidence in CA1 that interneurons participating in recurrent circuits, which exhibit

broad, complex and cell-specific spatial modulation (Kubie et al., 1990), are indeed tightly coupled with their presynaptic principal cells (Marshall et al., 2002; Maurer et al., 2006). Further, interneurons in both CA1 and CA3 have distinct separable pathways for recurrent synaptic transmission (for review, see Pelletier and Lacaille, 2008). Furthermore, this interpretation is bolstered by evidence in DG NR1 knock-out mice that NMDA-dependent plasticity in DG is necessary for rate remapping in CA3 but not the positional recoding in distinct contexts necessary for global remapping (McHugh et al., 2007). Behavioral and electrophysiogical evidence in rates further reinforce this decoupling of DG and CA3. Subregion-specific pharmacological blocking of NMDA receptors demonstrates behavioral differentiation, between the CA1/DG subregions and CA3, in the ability to transfer learned spatial tasks to new environments (Lee and Kesner, 2002). Further, general pharmacological block of NMDA receptors in a novel environment blocked the emergence of spatial selectivity in DG but left CA3 place coding unaffected (Leutgeb et al., 2007). This indicates that DG implements competition among its granule cells through processes of synaptic modification, whereas at least some of the capability of CA3 to form sparse codes may be dependent on intrinsic hard-wired network dynamics (Moser et al., 2008). If so, this may be unique among hippocampal subregions, as competitive selection of CA1 active subsets in novel environments is a long-timescale process likely to be plasticity-dependent (Karlsson and Frank, 2008). However, there is evidence suggesting that dynamic competitive selection of active units occurs in amygdala (Han et al., 2007), indicating that other brain areas may utilize network dynamics as a substrate for rapid memory formation.

As such, we suggest that this second interpretation both better matches the simple structure of our network model and allows for an interesting hypothesis of a parallel microcircuit in CA3 that makes prototype spatial codes immediately

available based on direct entorhinal input. Indeed, as DG and CA1 may require experience-dependent plasticity for competitive sparse coding, there may be a functional niche in hippocampal processing for a rapid "bootstrap" of a neuronal index based on direct cortical input. Indeed, such an index has been theorized for decades as a primary hippocampal function in service of episodic memory formation in neocortex (Teyler and DiScenna, 1986; Teyler and Rudy, 2007). In novel environments, this is particularly important because there is no previously learned spatial representation that can be autoassociatively retrieved (Marr, 1971). Furthermore, this hypothesis posits a role for the entorhinal layer II projection to CA3 along the perforant path (Witter, 2007) to drive this prototyping process. This is consistent with the idea that an active perforant path input is necessary to enable associative synaptic modification during the encoding of new spatial information (Treves and Rolls, 1992).

This role for inhibition may be a component of its overall information processing contribution to spatial learning and stabilization in the hippocampus (cf. Paulsen and Moser, 1998). Learning mechanisms in the DG and CA3 networks may be able to improve the spatial coding deficiencies of these prototype maps with experience. Several model show that long-term heterosynaptic depression may improve spatial specificity by de-tuning the afferent inputs contributing to secondary place fields (Rolls et al., 2006; Savelli and Knierim, 2007; Molter and Yamaguchi, 2008), though that is potentially limited by those that are activated by grid inputs that also contribute to a unit's primary place field. Furthermore, it has been observed that initially large active subsets in CA1 become smaller, in an apparently competitive manner, with long-timescale experience in a new environment (Karlsson and Frank, 2008). Associative synaptic modification in the DG and CA3 networks may similarly serve to competitively improve network spar-

sity, though the dynamics of plasticity and inhibitory control are likely to reflect mechanisms distinct from CA1 (Leutgeb et al., 2004; Nitz and McNaughton, 2004; Karlsson and Frank, 2008). Lastly, both models and experiments show that place fields can shift and skew (Gerstner and Abbott, 1997; Mehta et al., 1997) and become substantially larger with experience (Wallenstein and Hasselmo, 1997; Mehta et al., 2000; Rolls et al., 2006). These experience-dependent effects are likely to depend on excitatory associational synapses in the place population, which were not modeled here.

Chapter 4

Rapid Global Remapping Induced by Entorhinal Realignment

"No, of course you didn't go to Mars; you would know that, I would think. Aren't you always bleating about going?"

He said, "By God, I think I went." After a pause he added, "And simultaneously I think I didn't go."

"Make up your mind."

"How can I?" He gestured. "I have both memory-tracks grafted inside my head; one is real and one isn't but I can't tell which is which."

—Philip K. Dick, *We Can Remember It For You Wholesale*

4.1 Introduction

4.1.1 Hippocampal remapping

Since the discovery (O'Keefe and Dostrovsky, 1971) and early characterization (Ranck, 1973; O'Keefe, 1976; Muller et al., 1987) of the spatially selective responses of place cells in rat hippocampus (for review, see Section 1.3), the study of how spatial representations change has become a critical line of research into hippocampal function. Changes in environmental, contextual, motivational and other sorts of inputs can induce varying degrees of associated changes in hippocampal spatial representations. This is referred to generally as remapping.

Early on, remapping was explored by Muller and Kubie (1987), who examined place coding changes in response to minimal environmental modifications. Muller and Kubie found that rotations of proximal cues in a cylindrical enclosure elicited corresponding rotations of place fields, and that changing the scale of the enclosure caused about one-third of place cell responses to scale similarly. Such cue-dependence contributes to the coherence of spatial representations despite minor changes, but does not effectively restructure of the spatial code. However, over half of the cells they recorded in the rescaled cylinder did not also rescale, and instead developed place fields completely unrelated to their responses in the original environment. Quirk et al. (1990) then found that almost all place cells also remapped in the same environment depending on whether or not they were introduced to the chamber in darkness. This "complete remapping" demonstrated that place cells can support multiple independent population spatial maps. To emphasize that substantial changes in place-cell response could be driven by minor cue changes, Bostock et al. (1991) showed the development of divergent spatial maps over time by simply changing the color of a proximal cue card from white to black.

Subsequent studies of hippocampal remapping significantly expanded understanding of the causes, flexibility and diversity of remapping effects. There are distance-dependent effects of cue changes in conditions of rate remapping (Hetherington and Shapiro, 1997). Changing the color of the enclosure causes complete remapping (Kentros et al., 1998). Large mismatches in head-direction cues, greater than 45°, can provoke the instantaneous formation of a new representation (Knierim et al., 1998). Motivational changes in the animal, due to task requirements or conditioning, have been shown to elicit remapping: partial remapping and place field directionality can be task-dependent (Markus et al., 1995), different within-task reference frames may have independent maps (Jackson and Redish, 2007), and contextual fear conditioning can induce remapping (Moita et al., 2004). Further, different degrees of partial remapping can be evoked by moving a maze through three-dimensional space (Knierim and McNaughton, 2001) or rotating it through different reference frames (Zinyuk et al., 2000; Cressant et al., 2002). Notably, the extent of remapping may be limited to local sensory environments, such that connected spaces are independently represented (Paz-Villagran et al., 2004).

More recent studies have attempted to elucidate the subregional differences in remapping as a response to contextual change. Leutgeb et al. (2004) showed that subregion CA3 forms distinct cell assemblies in different rooms, regardless of enclosure similarity, whereas CA1 spatial codes were modulated by the similarity of the proximal environment. This was confirmed, along with demonstrating that CA3 active ensembles change less than CA1 ensembles for very minor alterations, by immediate-early gene imaging (Vazdarjanova and Guzowski, 2004). Then, Leutgeb et al. (2005) used population vector analysis of spatial correlation to argue for a primary dichotomy in remapping: that there is rate remapping, in which population rates randomly recode due to local context changes; and "global

remapping", in which both locations and population rates randomly recode due to a major shift in environment (e.g., a different room). Subsequent population vector studies of the transition between two distinct environments have shown both attractor-like dynamics and incremental adjustments commensurate with cue changes depending on whether the training protocol was shuffled (Wills et al., 2005) or progressive (Leutgeb et al., 2005), respectively.

4.1.2 Entorhinal realignment

After the discovery of grid cells in MEC (Hafting et al., 2005; Witter and Moser, 2006), the role of spatial information in entorhinal cortex with respect to hippocampal remapping became a critical question. Fyhn et al. (2007) performed simultaneous recordings of CA3 place cells and MEC grid cells under various remapping protocols and found that grid-cell responses changed contiguously with hippocampal remapping. This was observed in two global remapping conditions: a different enclosure in the same room caused a randomization of grid spatial phases; an identical enclosure in a different room produced random displacements of both the spatial phase and orientation of the grids. This grid "realignment" did not occur in conjunction with hippocampal rate remapping. Further, Fyhn et al. found that the colocalized ensembles of grid cells from which they recorded realigned coherently (i.e., by the same phase and orientation change). Notably, more recent preliminary data of simultaneous recordings from both hemispheres suggests that coherence in MEC realignment may extend beyond local ensembles (Hafting et al., 2008).

The observed contiguity of MEC realignment and hippocampal global remapping (Fyhn et al., 2007) is both temporal, in that they occur simultaneously and at the same short timescale, and of degree, in that they are both all-or-none phe-

nomena. Both components of this correspondence suggest a direct mechanism in which entorhinal response changes drive the formation of statistically independent hippocampal spatial codes in conditions that induce global remapping. The local coherence (Fyhn et al., 2007) and, preliminarily, the bi-hemispheric coherence (Hafting et al., 2008) of entorhinal realignment indicate the possibility that realignment is globally coherent across MEC. It is unclear how a global, coordinated shift in MEC response properties could produce divergent representations in hippocampus. Colgin et al. (2008) suggest two hypotheses: that there are multiple, independently realigning modules in MEC and that MEC is a large continuous map of space which re-references itself in different environments to randomize the spatial phases of grids with different spatial frequencies.

4.1.3 Modeling MEC realignment and remapping

Despite the overall diversity of remapping behavior, we will use our model of hard-wired spatial map formation (Chapter 3) to benchmark global remapping by simulating MEC realignment. That is, we restrict the approach to whether certain forms of randomization of cortical responses can effectively produce randomizations of the place and rate codes of the spatial maps. The basis for this assessment is a simulated remapping "experiment" involving the comparison of two environments: an initial reference environment A, determined by a randomly sampled phase and orientation vectors; and a modified environment B, created by adjusting the cortical representation of A in a way specified by the parameters of realignment. The model does not account for visual landmarks, path integration inputs, or other factors that affect place cell activity in hippocampus, all of which may have some role in remapping. However, the model we present is a functional readout of MEC grid firing rate maps, so our assessment is naturally restricted to rate-based

propagation of changes in entorhinal responses associated with the "orthogonalization" of spatial coding under global remapping (Leutgeb et al., 2005; Fyhn et al., 2007). Other models, for instance, demonstrate that temporal phase-based coding may contribute to place field formation (Molter and Yamaguchi, 2008), and thus may indicate another component of realignment-based remapping.

The detailed competitive balance for place units in our model is heterogeneous throughout the environment. This heterogeneity is a result of a different cortical input profile at every location in the environment. Combined with nonlinear integration of those inputs in each place unit, small input and place coding changes can be amplified. This mechanism may provide a form of activity-dependent pattern separation (Lee et al., 2004; Leutgeb et al., 2007), which can contribute to and enhance remapping despite local coherence in cortical response changes (Fyhn et al., 2007). Thus, we continue now to explore hippocampal remapping under various conditions of MEC realignment and suggest that modularity in realignment may be necessary to account for the complete orthogonalization of spatial coding inherent to global remapping. We use the hypothesis of modular realignment and the separability of phase and orientation adjustments to construct different realignment scenarios.

4.2 Methods

The same phenomenological model of MEC grid-cell responses and dynamic competitive model of spatial map formation as previously described (Section 3.3) are used here to assess entorhinal realignment and hippocampal remapping.

4.2.1 Simulation of MEC realignment

To explore remapping as a function of cortical realignment, we examine various realignment scenarios. Realignment is implemented by adding translational and angular adjustments to the spatial phase and orientation vectors, respectively, of the simulated MEC inputs. Unless otherwise specified, these realignment parameters are randomly sampled: phase adjustments uniformly from an annulus of 5.0–25.5 cm; orientation adjustments uniformly from 66 degrees clockwise (CW) to 78 degrees counter-clockwise (CCW). These ranges correspond to the observed range of MEC realignment (Fyhn et al., 2007). When phase and orientation realignment are combined, the rotation for the angular adjustment always follows the spatial phase adjustment and is then performed with respect to the mid-point of the 1m square environment.

Realignment modules are implemented as random equal-sized partitions of the simulated MEC population. In modular realignment scenarios, each module has its own independently sampled realignment parameters. For the realignment variance condition (Figure 4.11b), white noise vectors of the specified SD are added to the alignment vectors. In describing certain remapping examples in Section 4.3.2, will we refer to conditions of full phase and orientation realignment based on the number of modules: $N = 1$ indicates global coherence, whereas $N = 2$ and $N = 4$ indicate two and four independent modules, respectively.

4.2.2 Population measures of remapping

We measure remapping strength as a function of the pair-wise spatial coding structure of place units active in both an initial environment A and a modified environment B. For place (positional) remapping, we compute $1 -$ Pearson r of the

pair-wise distances between maximal response locations of units between environments. That is, if N place units are active in both A and B, then there are $n = N(N-1)/2$ pairs of coactive units. We compute inter-unit distance vectors $\vec{D}_A$ and $\vec{D}_B$ for each environment:

$$\begin{aligned} \vec{D}_E &= [D_E^p]_{p=1}^n \\ D_E^p &= \sqrt{(X_E^i - X_E^j)^2 + (Y_E^i - Y_E^j)^2}, \text{ for } (i,j) = k(p) \end{aligned} \tag{4.1}$$

where E is environment A or B, X_E^i and Y_E^i denote the maximal response location of unit i in environment E, and k is a list, indexed by p, of all coactive unit pairs (i,j). Then, place remapping strength is computed as

$$\begin{aligned} \Delta\text{Place(A,B)} &= 1 - \text{corr}(\vec{D}_A, \vec{D}_B) \\ &= 1 - \frac{n\sum D_A^p D_B^p - \sum D_A^p \sum D_B^p}{\sqrt{n\sum (D_A^p)^2 - (\sum D_A^p)^2}\,\sqrt{n\sum (D_B^p)^2 - (\sum D_B^p)^2}} \end{aligned} \tag{4.2}$$

where the sums are taken over all pairs $p \in \{1, n\}$. Similarly, for rate remapping, we compute $1 -$ Pearson r of pair-wise normalized rate differences between environments. Using the vectors $\Delta\vec{R}_A$ and $\Delta\vec{R}_B$,

$$\begin{aligned} \Delta\vec{R}_E &= [\Delta R_E^p]_{p=1}^n \\ \Delta R_E^p &= (R_E^i - R_E^j)/(R_E^i + R_E^j), \text{ for } (i,j) = k(p) \end{aligned} \tag{4.3}$$

where R_E^i is the maximum rate of unit i in environment E, we compute rate remapping strength as

$$\begin{aligned} \Delta\text{Rate(A,B)} &= 1 - \text{corr}(\Delta\vec{R}_A, \Delta\vec{R}_B) \\ &= 1 - \frac{n\sum \Delta R_A^p \Delta R_B^p - \sum \Delta R_A^p \sum \Delta R_B^p}{\sqrt{n\sum (\Delta R_A^p)^2 - (\sum \Delta R_A^p)^2}\,\sqrt{n\sum (\Delta R_B^p)^2 - (\sum \Delta R_B^p)^2}}. \end{aligned} \tag{4.4}$$

Values for ΔPlace(A,B) and ΔRate(A,B) range from 0.0, indicating no change in the relative structure of the place or rate code, to 1.0 (or possibly greater, due to

anti-correlation), indicating statistical independence between the spatial maps for environments A and B.

Single-unit remapping measures are also computed under various realignment conditions. As a measure of positional recoding, we computed the distance between maximal response locations in environments A and B (Figure 4.4a),

$$\Delta X_i = \sqrt{(X_A^i - X_B^i)^2 + (Y_A^i - Y_B^i)^2}. \tag{4.5}$$

Similarly, as a measure of rate recoding, we computed the normalized rate change for a single unit between environments (Figure 4.4b),

$$\Delta R_i = \frac{\max(R_A^i, R_B^i) - \min(R_A^i, R_B^i)}{\max(R_A^i, R_B^i)}. \tag{4.6}$$

This yields an unsigned value from 0.0–1.0 indicating the relative degree of rate change for unit i between A and B.

4.2.3 Realignment visualization

Realignment parameter sweeps (Figure 4.11) were created similarly to the parameter sweeps for spatial coding (Section 3.3.5). Here, a single random network fixed at the parameter reference-point of $J_{\text{inh}} = 2.5$ and $\lambda = 1.5$ was used. Randomly chosen spatial phase and orientation alignment vectors were used to create the cortical representation of the initial environment A, followed by simulating its corresponding spatial map. These alignment vectors were then modified as specified for the realignment sweeps and a spatial map for the the resulting modified environment B was simulated. For each sampled increment in the realignment adjustments, positional and rate remapping strengths (Section 4.2.2) relative to environment A were computed to visualize the realignment-dependence of remapping.

To visualize the unit pair-wise data underlying the remapping measures (Equations 4.2 and 4.4), we first computed the corresponding distance ($\vec{D}$) and rate dif-

ference ($\Delta\vec{R}$) vectors for all coactive pairs of place units. We then created two-dimensional histograms of the scatter data for positional ($\vec{D}_A$, $\vec{D}_B$) and rate ($\Delta\vec{R}_A$, $\Delta\vec{R}_B$) remapping for several example remapping experiments (Figure 4.6). Histograms consist of 64×64 bins representing ranges of up to 120 cm distances or $\pm$70% normalized rate differences.

Population cross-correlograms for remapped representations were computed in the same way as autocorrelograms (Section 3.3.5) except that the complex conjugate of the spatial ratemap for environment B was used for the convolution.

To visualize the statistics of remapping strength across various realignment conditions, we computed means and 95% confidence intervals (CIs) for both place and rate remapping strength for sample sets of 15 random remapping experiments (Figure 4.12). For each sample set, we plot an ellipse centered at the mean remapping strength whose width and height correspond to the place and rate remapping confidence intervals, respectively.

4.3 Results

We begin by presenting an example remapping experiment to show that random recoding occurs in individual place units. This is followed by a more comprehensive analysis of a series of examples showing different types of realignment. We use these examples to introduce strict pair-wise measures of place and rate remapping. Finally, we then proceed to systematic evaluations of the role of cortical non-coherence and the ability of entorhinal realignment to effectively drive global remapping.

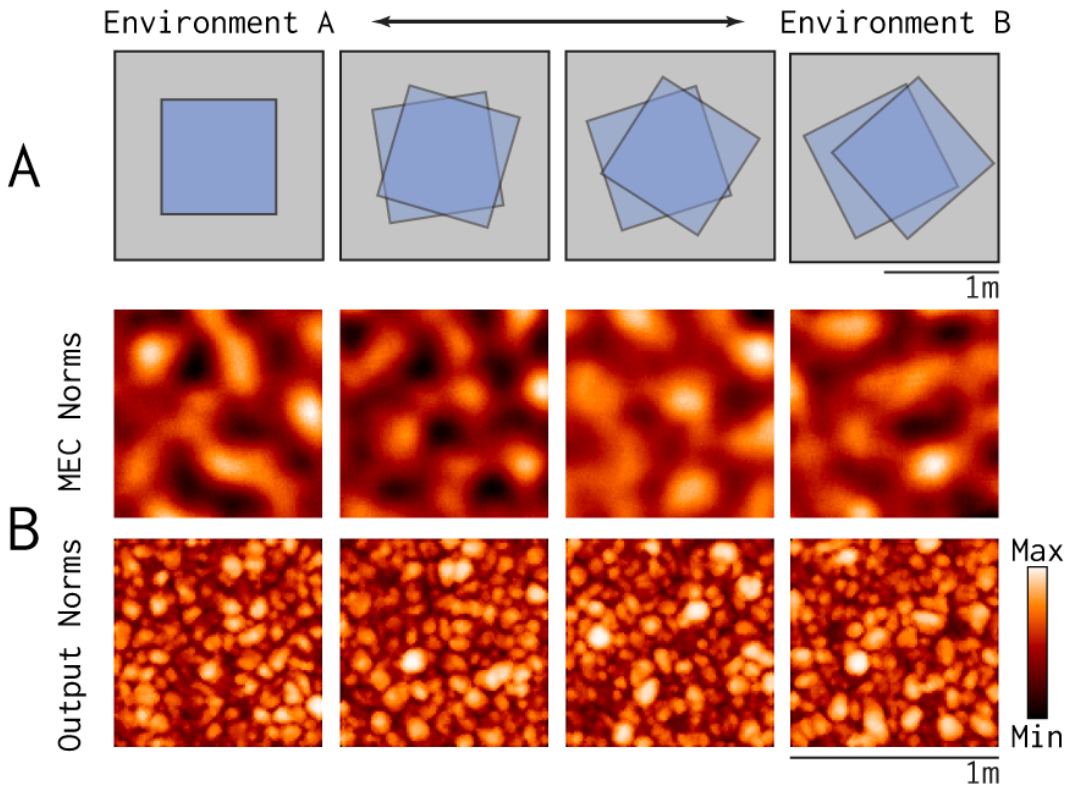

Figure 4.1 Example remapping experiment involving two independent realignment modules. A. Schematic of the two realignment modules as they adjust independently. B. The distribution of input fluctuations across the environment changes continously throughout MEC realignment. Each map is normalized to its respective maximum and minimum values to show the relative fluctuations in high contrast. Contiguously, the norms of the output spatial maps also change with realignment, in a way that is largely uncorrelated with concurrent cortical input norms. The output maps maintain sparse and informative spatial representations throughout.

4.3.1 Recoding with two modules

First, we present an example remapping experiment driven by two independently realigning cortical modules. The MEC population is randomly partitioned to create modules (see Methods), so each module consists of 250 grid maps in this example. Each module is assigned randomly sampled realignment parameters that specify the adjustments that will be added to the spatial phase and orientation vectors of the grid maps. In this way, the cortical representation of a random, initial

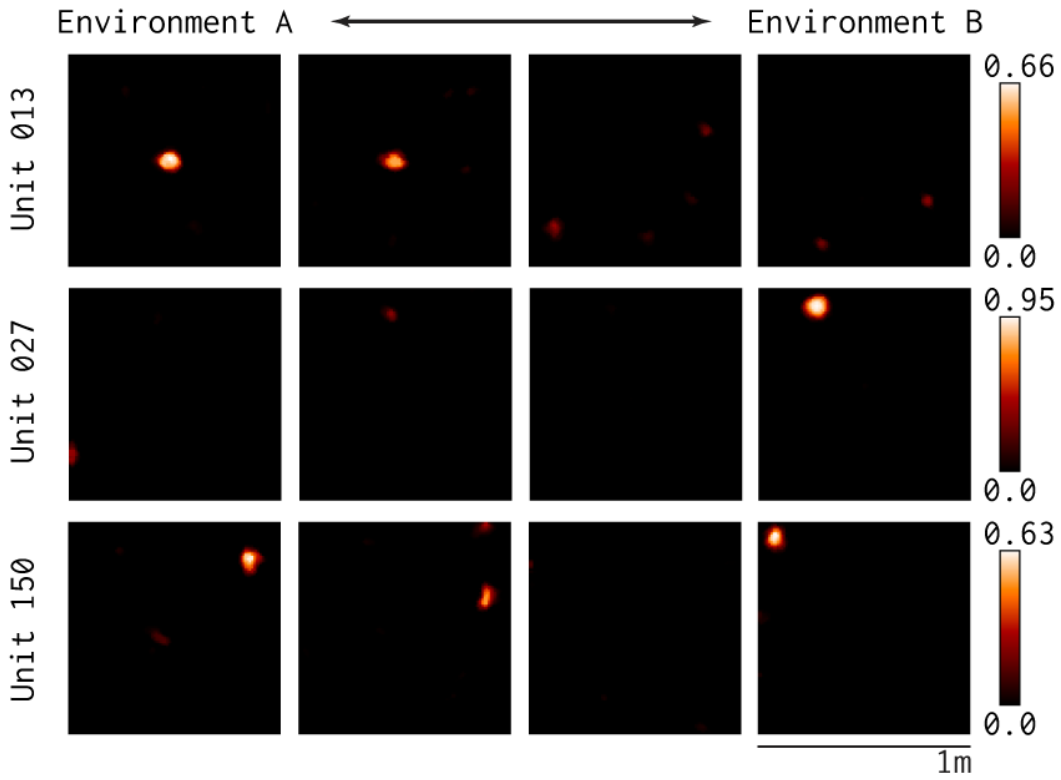

Figure 4.2 Place units recoding under MEC realignment. Individual response rate-maps of example place units demonstrate remapping contiguous with realignment of the cortical inputs.

environment A is modified to create the representation for a subsequent environment B. Simulated spatial maps for A and B can then be compared to assess the quality and degree of remapping that has occured in the outputs. Schematics of the two realigning modules are shown with two intermediate realignments (at 33% and 66% of full realignment) to illustrate the progression of cortical non-coherence (Figure 4.1a). Recall that we measure the magnitude of population rates by computing vector norms across the environment (Equation 3.6). MEC population activity (Figure 4.1b, top) and the corresponding spatial map produced by the model (Figure 4.1b, bottom) show that the distributions of input and output fluctuations change substantially as the realignment progresses. The output norms do not obviously follow changes to the distribution of input norms (Figure 3.14; Section 3.4.5).

Importantly, the output spatial code maintains a sparse but stable representation of the environment thoughout the realignment. There are no apparent large-scale changes to the quality of the spatial representation.

If these changes in the output norms are reflective of remapping, then individual place units should demonstrate a diversity of random recoding across the transition from environment A to B. Response ratemaps for three example active units show the full typology of recoding that would be expected of global remapping (Figure 4.2). Unit 13 loses a strong place field in A to become a dead unit in B. Unit 27 does the opposite, responding with a strong field only as the realignment to environment B completes. Unit 150 recodes by forming a place field in B in a location distinct from its similarly strong field in A. Notably, units 13 and 150 demonstrate responses in the 33% realignment that are similar but weakened and/or shifted compared to their initial place fields in environment A.

4.3.2 Remapping under different types of realignment

We now examine five example remapping experiments driven by different MEC realignment conditions (Figures 4.3–4.10): coherent rotation (45° CCW; 'rotate'); coherent shift ($(\Delta x, \Delta y) = (11.0, -8.0)$ cm; 'shift'); full coherent realignment (combined rotate and shift; 'N = 1'); and full modular realignment (randomly sampled parameters[1]; 'N = 2' and 'N = 4'). The realignment parameters for the three coherent conditions were chosen as mid-range values to represent what may be 'typical' realignment. Below, the examples will be referred to by the labels for their respective realignment conditions.

[1]Modular realignment parameters: N = 2: [7.4° CCW, $(-2.9, 15.4)$], [16.2° CCW, $(-19.4, 13.6)$]; N = 4: [48.9° CW, $(13.9, -2.6)$], [16.2° CW, $(-20.7, -10.2)$], [26.7° CW, $(-12.6, 17.4)$], [56.2° CCW, $(-14.0, -16.5)$]

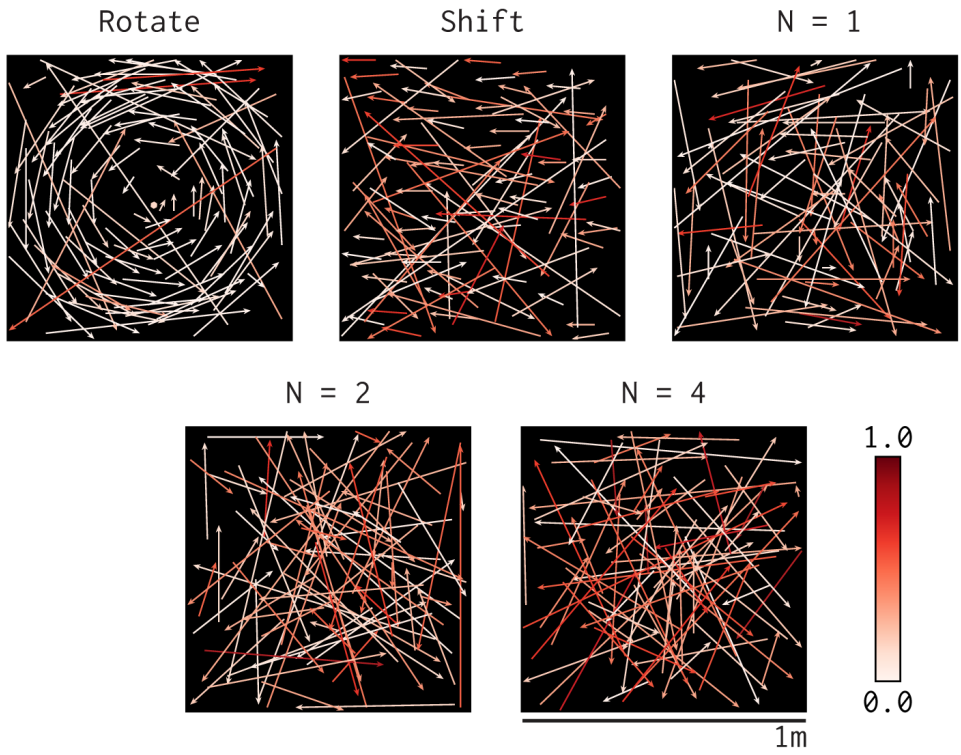

Figure 4.3 Effective place and rate remapping is only apparent when grid rotation is combined with phase shifts. Examples are shown of random networks remapping from novel environment A to novel environment B under different realignment conditions: orientation realignment only (rotate); phase realignment only (shift); and full realignment with population coherence (N = 1), two coherent modules (N = 2), and four coherent modules (N = 4). Recoding is shown for place units active in both environments A and B. Arrows for each unit show the change in maximal response location from environment A to B and are colored by normalized peak-rate change (white, no change; red, maximal change).

To directly visualize the recoding of the spatial map between environments A and B, we constructed quiver plots over the environment (Figure 4.3). All place units that are active in both environments are represented by an arrow, with the tails and heads positioned at the maximal response locations in A and B, respectively. Unfortunately, there is no satisfactory way to represent units that are active in only one environment, so they are not included. Rate recoding is measured as a normalized rate change (Equation 4.6) and represented by the color of the

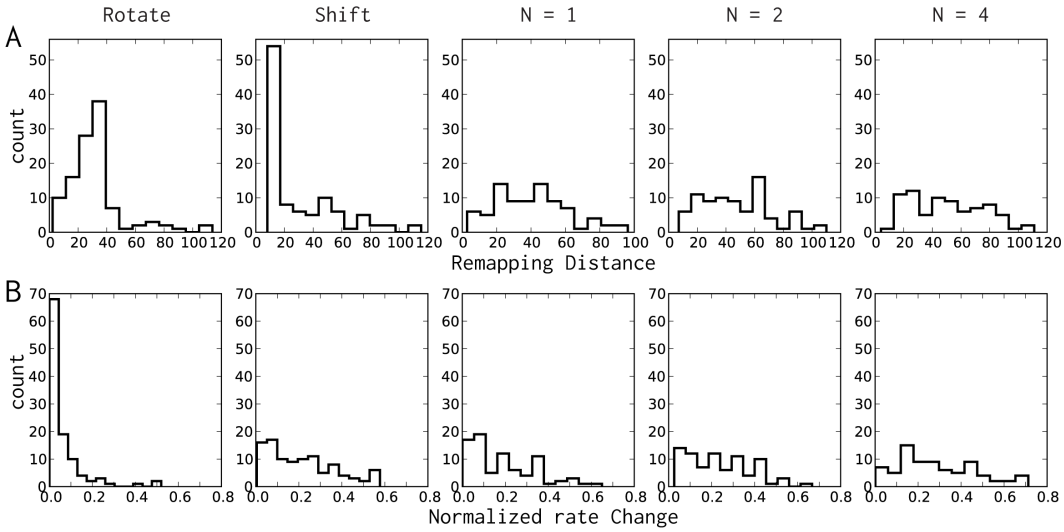

Figure 4.4 Distributions of peak positional and rate changes show the recoding behavior of individual place units. A. Distance between peak rate locations in environment A and environment B show a coherent remapping mode in the shift condition, and more dispersed mode for the rotate condition. B. In contrast, normalized peak rate changes between A and B show a coherent mode in the rotate condition that is not apparent in the other realignment conditions. More uniform recoding distributions are indicative of better remapping, but these per-unit measures are not dispositive for the complete orthogonalization inherent to global remapping.

arrow, with white indicating no change between environments and red indicating maximal change. The rotate condition, which is not observed experimentally, is a simple coordinate transform of the cortical inputs. Thus, only a small subset of units randomly remap, with most following the rotation of the inputs. The units that shifted position also tended to significantly change their peak response rates while those that followed the inputs did not. Histograms of these single-unit positional shifts (Equation 4.5, Figure 4.4a) and rate changes (Equation 4.6, Figure 4.4b) for the rotate condition further illustrate that few units recode substantially. The mode for positional recoding, though, is more dispersed than for rate changes since the distance that an active field in A shifts generally depends on its distance from the mid-point of the environment. However, the mode for

rate changes is somewhat dispersed, with at least 29 units beyond 0.0% showing rate changes up to 15%. This indicates that, even in this worst-case condition for eliciting remapping, the rate code is being altered by the small subset of units that randomly remap to different regions of the environment (i.e., those shifting >50 cm in Figure 4.4a), changing the local competitive balance within pre-existing neighborhoods of place fields. The shift condition, similarly, shows a division between place units that shift their responses coherently and those that randomly remap (Figure 4.3), though rate recoding appears more substantial across the population than in the rotate condition. This is verified in the histograms: the narrow mode around 13.6 cm for positional shifts is due to the input-following subpopulation while the long tail results from the random recoders (Figure 4.3a); further, there is no strong coherent mode for rate change, indicating that the population rate code was randomized (Figure 4.3b). Spatial phase shift, even completely coherent, enables more random recoding by causing more place fields to lose their respective cortical input peaks. The larger subset of units that remap is then able to decohere the rate code throughout the environment by upsetting the competitive balance between neighborhoods of input-following place units.

Global remapping in response to environmental change may also involve both components of MEC realignment (Fyhn et al., 2007). Combining the coherent rotate and shift conditions, referred to here as $N = 1$, eliminates any visually apparent coherence in the positional recoding of place unit between environments (Figure 4.3). The positional histogram for $N = 1$ (Figure 4.4a) does not have a coherent mode and is more symmetrically distributed across the range of distances from 0–120 cm. The rate-change histogram (Figure 4.4b) does not show a strong mode but is distributed similarly to that of the shift condition alone. So, from both the remapping quiver plots and histograms of per-unit recoding data, full MEC

realignment appears to induce what looks like global remapping of the place-unit population. The remapping plots for the two modular conditions, $N = 2$ and $N = 4$ (Figure 4.3, bottom row), appear more randomized than that of $N = 1$, but it is unclear from the recoding histograms if there is any difference (Figure 4.4). More informative quantifications of remapping are necessary.

Population ratemap cross-correlation is a measure of spatial response similarity that incorporates the total response of the spatial maps for environment A and B. As such, it may be more informative than the measures used above that are based on just the locations and peak rates of maximal responses. Population cross-correlograms (see Methods) are shown for each of the five realignment condition examples (Figure 4.5). The rotate condition shows an expanded central peak on the spatial scale of the entire environment. The shift condition shows an offset peak correlation at the same scale as the autocorrelation of spatial output (Figure 3.18), indicating little structural change to the spatial code. Full coherent realignment produces a more dispersed central peak similar to but weaker than the rotate condition. The modular conditions (Figure 4.5, bottom) yield significantly more dispersed and weaker correlations than the coherent case, except for what appears to be an offset peak in the $N = 2$ condition. As evidenced by the relative dispersion of the central peak correlations, realignment modularity decorrelates the spatial response more effectively than the coherent conditions.

4.3.3 Pair-wise measures of remapping

Now, as a measure of the change of place and rate codes, we examine the relative structure of relationships between active place units. That is, we want to assess whether subpopulations of place units maintain the same or similar pair-wise inter-unit distances (Equation 4.1) or rate relationships (Equation 4.3) across MEC

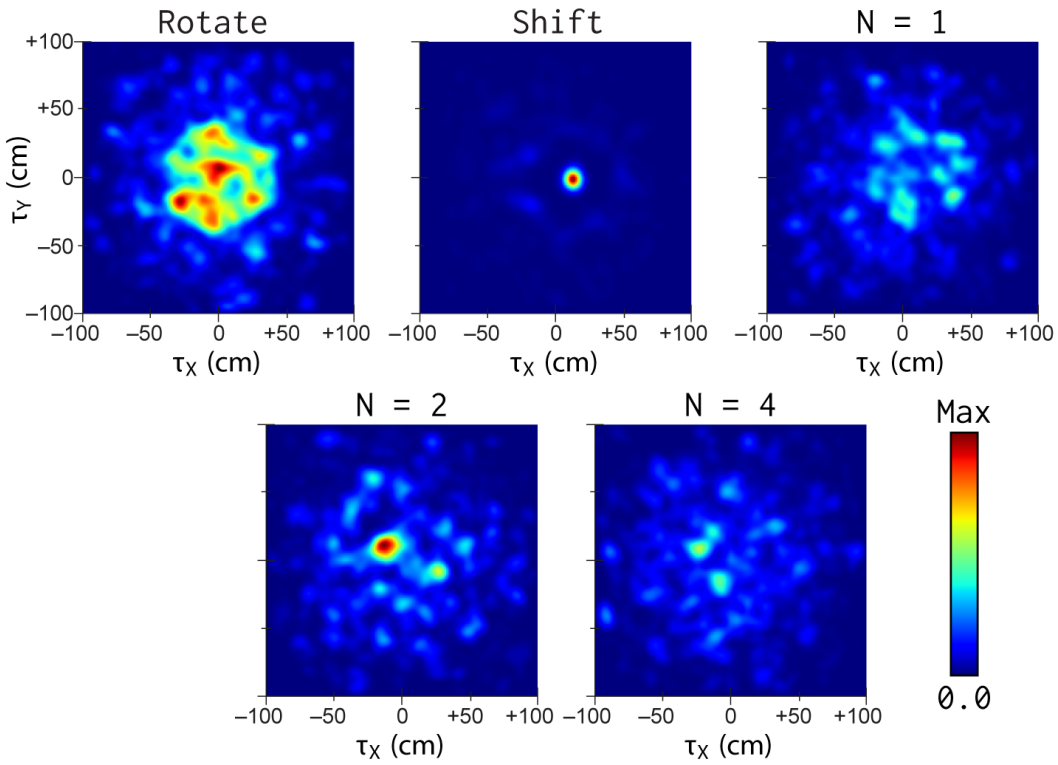

Figure 4.5 Population rate-map cross-correlograms for A–B remapping examples. Cross-correlograms are computed in the same way as autocorrelograms (see Methods) except that the spatial map of environment B, the result of MEC realignment from A, is complex conjugated for the Fourier-domain convolution. MEC rotation (rotate) produces significant dispersion of the central peak, which nonetheless indicates strong autocorrelation. Translation (shift) yields an offset central peak of the same scale as output autocorrelation (Figure 3.18b,c). Full realignment, combining shift and rotation, shows both significant dispersion and attenuation of centralized correlation across different modularity conditions. Rate-based decorrelation does not reveal substantial differences in remapping capabilities between non-modular ($N = 1$) and modular ($N = 2$, $N = 4$) MEC realignment. Each correlogram is normalized to the maximum value across conditions, so colors represent relative correlations.

realignment. Unike the per-unit measures and ratemap correlations discussed above, a strict population measure of structural remapping will not respond to coordinated shifts or other correlated changes among place units. Thus, for all pairs of units that are active in both environments A and B, we show inter-unit maximal

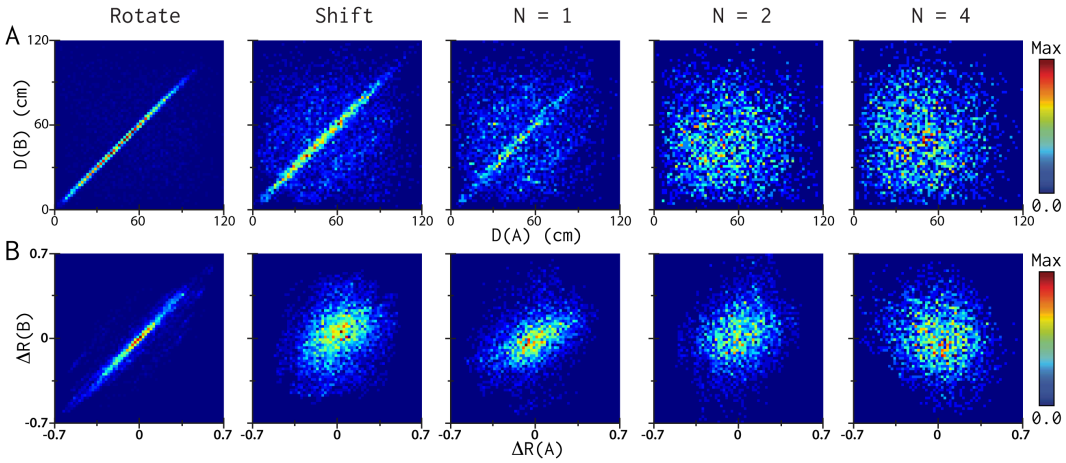

Figure 4.6 Pair-wise measures of A–B change reveal residual structure of previous spatial codes. Remapping shown as two-dimensional histograms (64×64 bins) for pair-wise positional (A) and rate-based (B) relationships between all units active in both environments. A. Distance between maximal response locations in A (D(A)) or B (D(B)) (Equation 4.1). Histogram maximum counts (left to right): 77, 20, 14, 8, and 7. B. Normalized difference between peak rates in A (ΔR(A)) or B (ΔR(B)) (see Equation 4.3). Histogram maximum counts (left to right): 79, 20, 17, 15, and 10. Only the modular realignment conditions fully randomize the pair-wise structure of the original spatial map.

response distances (Equation 4.1; Figure 4.6a) and rate differences (Equation 4.3; Figure 4.6b) as two-dimensional histograms for each of the realignment examples. These are histograms of the scatter data for environment A (x-axis) against B (y-axis), so any pair-wise structure of the place and rate codes retained across realignment will contribute to the positive diagonal.

Significant pair-wise correlations withstand realignment in the rotate condition for both the place and rate components of the spatial map. This is commensurate with the lack of structural change evident in the remapping diagram (Figure 4.3). Similarly, the shift condition shows a strong positive diagonal for the place code, but also a non-trivial amount of orthogonalized pair-wise recoding dispersed throughout the range of the histogram (Figure 4.6a). The input-following

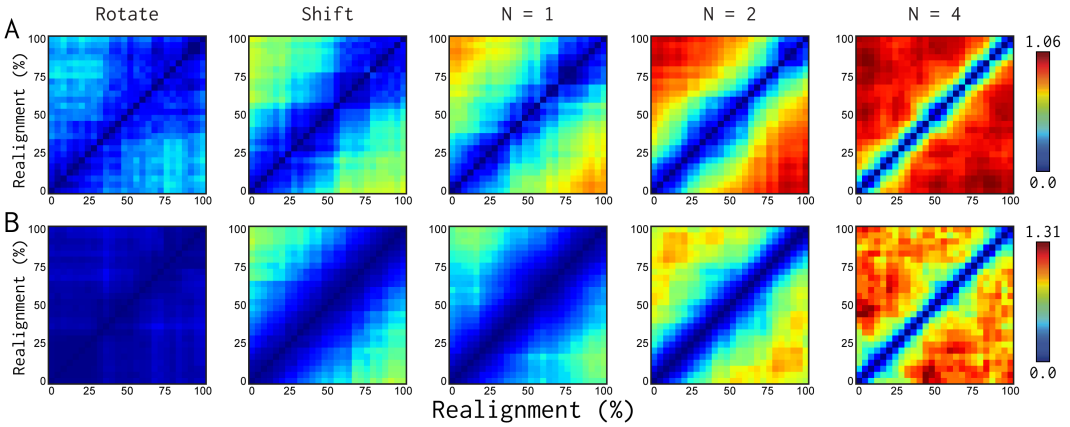

Figure 4.7 Pair-wise remapping strength across A–B realignment transition. Place (A) and rate (B) remapping strengths are shown between every pair of spatial maps as the MEC inputs realign from environment A to environment B. There are 25 intermediary spatial maps, so that each pixel here represents a 4% increment in realignment progress. Values $\geq$1.0 indicate complete orthogonalization of the realigned place and rate codes. The width of the diagonal bands is related to the relative effectiveness of a particular realignment condition to drive remapping, so that a narrow band indicates fast and efficient remapping.

place units, as a subpopulation, have maintained a coherent positional structure, contributing to the diagonal. The random recoding units yield randomized pair-wise distances. The rate histogram for the shift condition, however, is highly dispersed with no apparent diagonal correlations (Figure 4.6b). That is, the coherent shift condition produces both partial positional remapping and substantial disruption of the population rate code. The full coherent realignment example ($N = 1$) demonstrates similar positional remapping as the shift condition, except that the diagonal mode is weaker. Though, notably, $N = 1$ demonstrates more apparent rate-code correlation than the shift condition. Both of the modular conditions appear, at least visually, to have fully orthogonalized both the place and rate structure of the original spatial map. Thus, measuring pair-wise relative changes may be more informative than single-unit and population rate-based measures.

We proceed to quantify pair-wise remapping strength in order to assess the progression of remapping as environment A transitions into B. We can compute a scalar measure of place (Equation 4.2) and rate (Equation 4.4) remapping strength by taking $1-r$ for the respective pair-wise scatter data, where r is the Pearson correlation. These measures will allow us to quantitatively assess how effectively MEC realignment can drive remapping. For each realignment condition, we simulated 25 intermediate spatial maps corresponding to 4% increments in the realignment adjustments from the MEC representation of environment A to that of B. Realignment matrices are shown for the place (Figure 4.7a) and rate (Figure 4.7b) remapping strength between all of these spatial maps spanning A to B. The width of the diagonal band of these matrices indicates the proportion of total realignment over which remapping occurs, though the non-modular conditions (rotate, shift, and $N = 1$) fail to achieve full orthogonalization. As such, the diagonal width is a proxy for the overall effectiveness and, if realignment occurs over time, the relative time-scale of a particular realignment condition for driving the divergence of spatial codes.

Now, to show the convergence and extent of the remapping process as realignment progresses, we can examine just the place and rate remapping strengths with respect to environment A (Figure 4.8). The shift condition remaps gradually, equivalently for the place and rate codes, and falls short of global remapping. Full coherent realignment ($N = 1$) likewise falls short while exhibiting a relative deficit in rate recoding. Indeed, the $N = 1$ condition exhibited improved place remapping than shift (0.79 vs. 0.66) but worse rate remapping (0.60 vs. 0.72), which corroborates a visual comparison of the pair-wise data for the two example conditions (Figure 4.7). This is further verified by the statistics of remapping strength for random sample sets (Section 4.3.6). Both of the modular conditions, however,

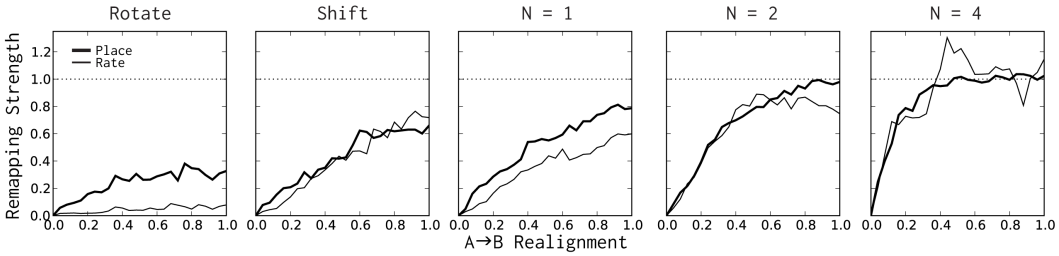

Figure 4.8 Progression of remapping as environment A gradually transitions into environment B. We simulated 25 intermediary spatial maps using linear interpolation of the spatial phase and orientation vectors that determine the alignment of the MEC inputs. Place- and rate-remapping strength are computed as pair-wise measures of the relative structure of the spatial code (Eqnations 4.2 and 4.4). Only the modular realignment conditions are able to completely orthogonalize the spatial representation for environment B relative to A; more modules does so faster and more effectively.

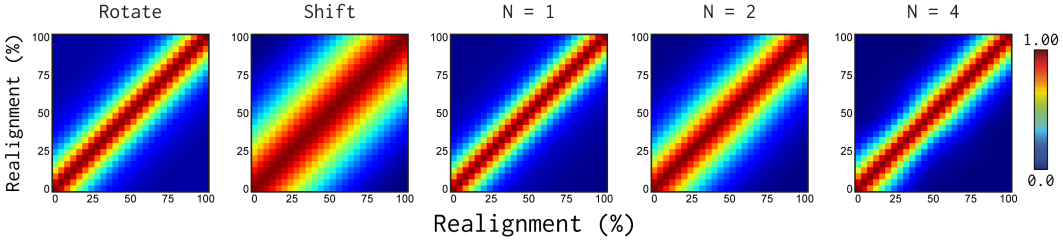

Figure 4.9 Population rate correlations throughout MEC realignment. Pair-wise realignment matrix shows that the decorrelation of absolute population rate vectors occurs smoothly as the cortical inputs realign. As with remapping strengths (Figure 4.7), a narrow diagonal band indicates relatively rapid decorrelation with respect to realignment.

orthogonalize place and rate codes by 80–90% (N = 2) and 30–40% (N = 4) of total realignment. Thus, fast and effective remapping may only be attained under conditions of modular realignment and a small number of modules is sufficient.

In the context of a realignment transition between two environments, we can examine the more traditional measure of population correlation. Here we measure the pixel-by-pixel correlations of the population ratemaps for every pair of intermediate spatial map in the realignment (Figure 4.9). The width of the diagonal is

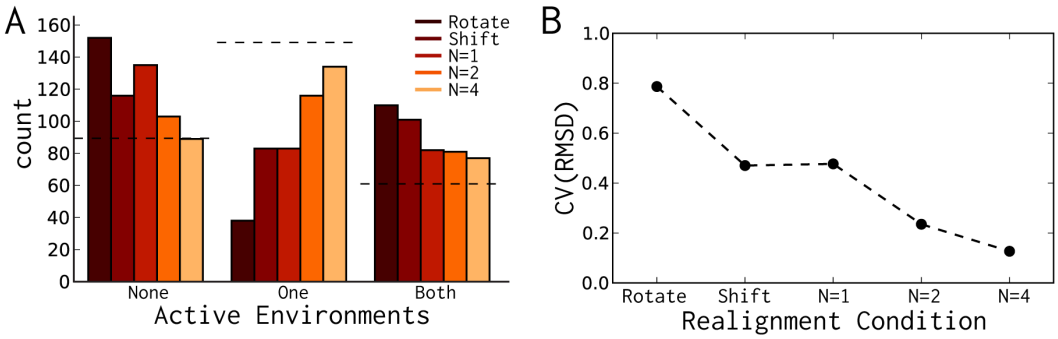

Figure 4.10 Recruitment of independent active subsets is enabled by modular MEC realignment. A. For each realignment condition, the number of place units active in none, one or both of the environments. Dashed lines show expected values if active subsets were independent across environments (assuming the average sparsity here of 54.8%). B. The mean-normalized root-mean-square deviation (CV(RMSD)) from the expected values indicates that only the modular realignment conditions (N = 2, N = 4) independently recruit active subsets.

consistent across the realignment and the decorrelation is smooth as realignment progreses. The diagonal band of the rotate condition is as narrow as the $N = 4$ band, even though former is a simple coordinate shift (Figure 4.3) while the latter yields complete orthogonalization (Figure 4.8). This demonstrates that population rate decorrelation is a poor measure of remapping; it is more dependent on the high-dimensional distance between the spatial maps for A and B than on structural changes in the spatial code. Our pair-wise measures of relative changes to the spatial map (Figures 4.7 and 4.8) can differentiate between poor remapping (rotate and shift conditions) and effective remapping (modular conditions).

4.3.4 Independent recruitment of active subsets

We can test whether active subsets between environments are independently recuited by considering the number of environments in which place units have active fields (Figure 4.10a). The realignment conditions that remap well tend to have

a larger proportion of place units that code for only a single environment, whereas poorly remapping conditions have more bipartite populations of coding and non-coding units. If we take the average network sparsity across these experiments (54.8%) and assume independent recruitment, then we would expect 90, 149, and 61 place units to be active in none, one, or both environments, respectively (dashed lines). The normalized root-mean-square deviation (CV(RMSD)) from these values for each condition, which varies from 0.79 (rotate) to 0.13 (N = 4), shows that only modular realignment enables approximately independent recruitment of active subsets (Figure 4.10b).

4.3.5 Realignment coherence

Since our example of coherent MEC realignment ($N = 1$) is able to produce substantial remapping, but not full orthogonalization (Figure 4.8), we examine the limits of the capability of coherent realignment to drive remapping. Fyhn et al. (2007) observed phase realignment (shift) in the range 5.5 to 27.5 cm and orientation realignment (rotation) in the range 66° CW to 78° CCW. The upper bound on shift roughly corresponds to half of the maximum grid spacing they observed. We simulated a realignment sweep (see Methods) of coherent shift from 0.0 to 25.0 cm and coherent rotation from 0° to 78° CCW (Figure 4.11a). For each of the samples, we computed measures of remapping strength that quantify changes to the relative pair-wise structure of place-unit maximal response locations and rates (see Methods). As such, these measures are not responsive to simple coordinate transforms or correlated changes across the population. The maximum realignment achieves 0.86 place and 0.84 rate remapping. A moderate realignment of 12.5 cm shift and 39° rotation achieves 0.74 place and 0.71 rate remapping. While this indicates substantial remapping, significant structural characteristics of the initial

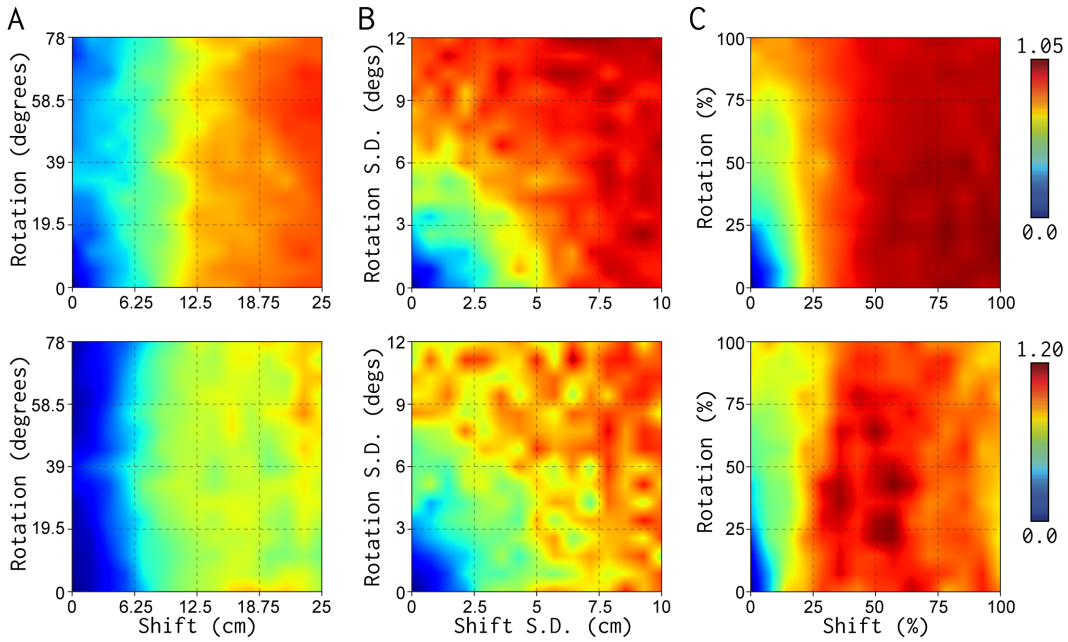

Figure 4.11 Non-coherent realignment of MEC grid-cell responses is required for orthogonalization of hippocampal spatial coding. Each panel shows positional (top) and rate (bottom) remapping strength for a realignment sweep of phase realignment (shift) varying with orientation realignment (rotation) under three example realignment scenarios (A–C). A. Population-wide realignment coherence: every grid map undergoes the same shift and rotation. The realignment bounds correspond to the maximum observed shift and rotation associated with global remapping (27.5 cm and 78° CCW; Fyhn et al., 2007). B. Total non-coherence in realignment: white noise is added to grid spatial phase and orientation across the MEC population. C. Modular realignment: the MEC population is divided into two independently realigning modules (see Methods). Spatial code orthogonalization is only available under conditions of non-coherent realignment.

spatial representation remain. As would be expected, rotation by itself has very little effect: with no shift, full rotation achieves just 0.19 place and 0.04 rate remapping. Further, rotation fails to have much effect even when combined with input shift across the range tested.

Next, we assess remapping under conditions of non-coherent MEC realignment. The most obvious way to introduce non-coherence is to assume no coher-

ence and have each grid map realign randomly and independently. We performed a realignment sweep injecting a range of additive white noise into the MEC spatial phase and orientation vectors. The SD of the white noise ranges from 0.0 to 10.0 cm for shift and 0 to 12° for rotation (Figure 4.11b). Non-coherence in both the shift and rotation of inputs is effective in leading to significant remapping: 10 cm SD noise in shift yields 0.95 place and 0.97 rate remapping; 12° SD noise in rotation yields 0.86 place and 0.72 rate remapping; combined, the spatial map undergoes 0.96 place and 1.0 rate remapping. Furthermore, small amounts of noise are sufficient to achieve remapping similar to the maximum realignment in the fully coherent condition: 5 cm shift SD and 6° rotation SD yield 0.85 place and 0.79 rate remapping. Thus, population variance among realigning MEC grids enables effective remapping. A realignment sweep for two such modules demonstrates that bisecting the MEC produces effective remapping long before the realignment is complete (Figure 4.11c). Here, both modules independently follow randomly sampled realignment parameters: the first module shifted by $(\Delta\, x, \Delta\, y) = (-23.8, 6.2)$ cm and rotated by 36.0° CW; the second shifted by $(5.8, -20.1)$ cm and rotated by 50.7° CW. Note that the random realignment shifts happen to be anti-correlated while the rotations are correlated; this corresponds well to the rapid remapping with shift and the more gradual remapping with rotation. Even so, rotation alone provides 0.82 place and 0.74 rate remapping; shift provides 1.0 place and 0.93 rate remapping; the full realignment yields 0.97 place and 0.79 rate remapping. Thus, modular realignment, even with just two modules, combines the effective remapping capability of population variance while preserving local coherence.

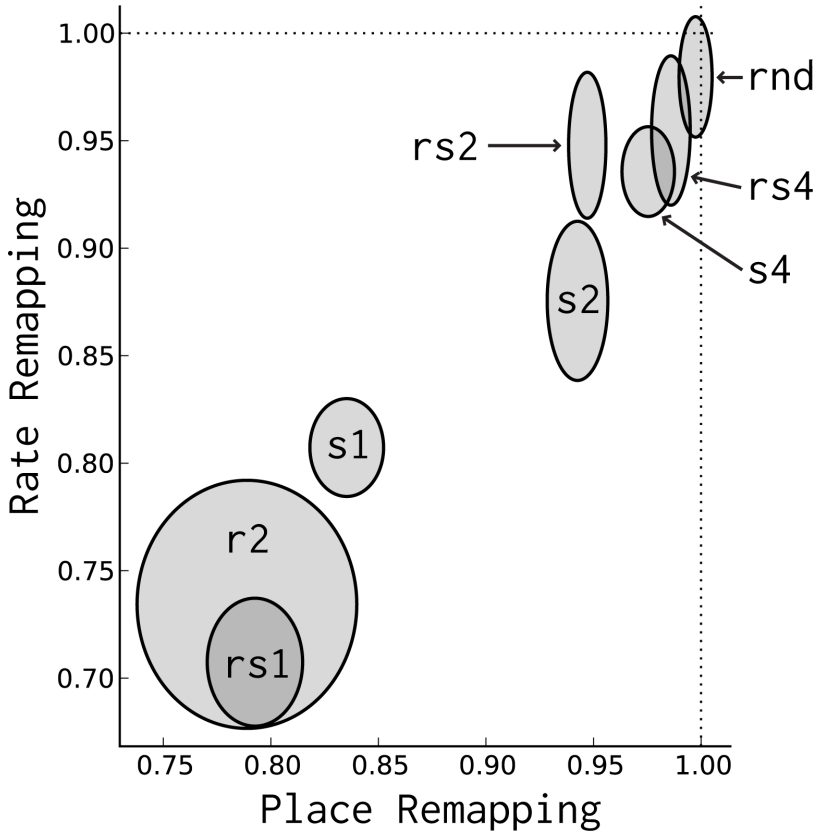

Figure 4.12 Remapping strength statistics for random sample sets of A–B experiments. A small number of realignment modules are sufficient to orthogonalize spatial coding. Means and 95% confidence ellipses (see Methods) are shown for sample sets of 15 remapping experiments under a variety of realignment conditions: labels indicate some combination of orientation realignment (r), phase realignment (s) and realignment modularity (1, 2 or 4). A set of random experiments, where both environments A and B were independently and randomly sampled, indicates the statistics of spatial code orthogonalization (rnd). The only sample set to intersect the random set is that of full realignment with four modules (rs4). Notably, orientation realignment only enhances remapping under conditions of modular realignment.

4.3.6 Statistics of remapping and MEC modularity

Finally, we consider the statistics of remapping strength for a variety of MEC realignment scenarios. For various combinations of shift, rotation, and modularity of N = 1 (full population coherence), N = 2 or N = 4, we ran 15 random remap-

ping experiments. For each experiment, a random network created a spatial map of a random environment A, which was then realigned according to randomly sampled realignment parameters to create a spatial map of environment B. The place and rate remapping strengths between the spatial maps for A and B for each realignment condition are shown as 95% confidence ellipses in Figure 4.12 (see Methods). The case of coherent rotation yielded only 0.33 ± 0.035 place and 0.13 ± 0.026 rate remapping (mean $\pm$ 95% CI; not shown). In line with the previous results for example remapping experiments, the full coherent realignment (rs1) condition fares worse than the coherent shift (s1) condition, though the difference is only significant for rate remapping (place: Student's $t = -1.44$, N.S.; rate: $t = -2.52$, two-tailed $p < 0.02$). However, with two modules, the rotation component of full realignment (rs2) enhances mean rate remapping over the shift condition (s2), but the difference is not significant (place: $t = 0.259$, N.S.; rate: $t = 1.36$, N.S.). Further, increasing modularity in full realignment to N = 4 significantly enhances place but not rate remapping (place: $t = 2.95$, $p < 0.01$; rate: $t = 0.134$, N.S.). As a benchmark for spatial code orthogonalization we ran a control set (rnd) in which environment B was a random environment unrelated to environment A. The sample set means and confidence ellipses show a progression of overall remapping from the shift condition (s4) to full realignment (rs4) to the random set (rnd) (Figure 4.12, upper right). These experimental sets are not significantly different from the control condition, though rs4 (place: $p = 0.37$; rate: $p = 0.60$) is more similar to the rnd set than s4 (place: $p = 0.16$; rate: $p = 0.24$). For MEC realignment and our model of spatial representation, four modules is minimally sufficient to create statistically independent spatial maps for novel environments that are structurally unrelated to a previous spatial map.

4.4 Discussion

4.4.1 Realignment coherence and remapping

The hypothesis of MEC realignment as the immediate cause of hippocampal remapping does not appear consistent with evidence showing that realignment is coherent both within colocalized ensembles (Fyhn et al., 2007) and, preliminarily, bihemispherically (Hafting et al., 2008). If realignment is globally coherent, as these data suggest, then it becomes difficult to see how any readout of MEC responses could produce divergent outputs. Thus, some source of global non-coherence may be necessary to account for global remapping. Two hypotheses have been proposed, initially by Fyhn et al. (2007) and then elaborated upon by Colgin et al. (2008): first, that there are distinct modules in entorhinal cortex that realign independently; second, that MEC represents a large continuous map of space which is re-referenced in different environments to randomize the spatial phases of grid-cell responses. These possible realignment mechanisms influenced our assessment of remapping here.

The continuous-map hypothesis entails a large absolute map independent of individual spatial representations that must be maintained across long timescales. Hypothetically, novel environments would provide a new random index into the continuous map, while familiar environments would induce the retrieval of an index that restores the original reference frame. Sensory-dependent place–grid association could provide a mechanism for such index retrieval and stabilization (O'Keefe and Burgess, 2005). Regardless, this absolute, indexable, long-term map would have to be accounted for by models of grid cell activity, which are typically based on either attractor dynamics or oscillatory interference (for review, see Giocomo and Hasselmo, 2008). However, there is evidence for both func-

tional and anatomical modularity in entorhinal cortex (Walling et al., 2006; Witter and Moser, 2006), similar in scale to isocortical columns. If that is the case, then independent modular realignment may be the more direct path to global non-coherence. Though, the effective difference between the two hypotheses is small: the continuous-map hypothesis entails that rotation must be globally coherent and that spatial phase randomization depends on the spatial frequency of grid cells. Indeed, Barry et al. (2007) found that grid-cell frequencies tend to cluster around fixed, non-integer ratios of some minimum scale (cf. Kjelstrup et al., 2008). Such clustering could effectively provide realignment modularization based on spatial frequency, with the number of modules scaling with the degree of clustering. Since coherent rotation does not significantly contribute to our pair-wise measurements of remapping, the s4 condition that we tested (Figure 4.12) would be equivalent to simulating the continuous-map mechanism given four frequency-based modules. The s4 condition remaps very effectively but does not completely orthogonalize, so a putative continuous-map mechanism might depend on more fine-grained clustering of grid spatial frequencies. For our purposes, the distinction between random modules and frequency modules is irrelevant, as we are only concerned with overall remapping capabilities and not scale-dependent effects.

In comparing the coherent with the non-coherent realignment conditions, we can consider the role of the rotational component of realignment. Rotation of grid-cell responses was only observed when an animal was remapping between two different rooms (Fyhn et al., 2007), as opposed to different enclosures in the same room. Two interpretations are possible: first, that rotation is a meaningful result of a more substantial remapping process caused by being in a physically different location; or, second, that rotation reflects the lack of a common reference frame between the two rooms for the inputs that determine grid orientation. Here, we

treated rotational realignment as a separable and potentially important constituent of the response that generates global remapping. We found that rotation alone, as expected, does not contribute significantly to remapping. In coherent realignment, the addition of rotation is actually detrimental to both place and rate remapping (see below). However, when considering rate decorrelation as the measure of remapping, the combination may remap more efficiently (Figure 4.9). In modular realignment, adding rotational adjustments does enhance remapping overall but not substantially. The vast proportion of the random recoding of spatial maps between environments is the result of the spatial phase displacements. Noting these observations does not make either of the interpretations of rotation more likely, but we suggest that effective realignment-based remapping should be considered a function primarily of spatial phase and not orientation.

4.4.2 Network dynamics as a basis for remapping

One of the striking features of the realignment of grid-cell responses in MEC is its contiguity with global remapping in CA3, evidenced not only by their all-or-none correspondence but also the nearly instantaneous transition in spatial correlations (Fyhn et al., 2007). This observation motivated our exploration of a model of spatial map formation constrained by hard-wired network dynamics: if such a model could create informative spatial representations on the timescale of inhibitory feedback, then it could also provide a basis for fast transitions when MEC responses decorrelate in distinct environments. It has been shown that CA3 rapidly incorporates new spatial information (Nakazawa et al., 2003; Lee et al., 2004; Leutgeb et al., 2006; Leutgeb and Leutgeb, 2007), but that correlations stabilize on a longer timescale than CA1 (Leutgeb et al., 2004). This prolonged stabilization may be the result of plasticity among the dense recurrent excitatory synapses of CA3 pyra-

midal cells, but also perhaps due to the reconciliation of spatially informative input from DG, as it becomes available, with the direct-input prototype map (Section 3.5.3). Regardless, the relative duration of the evolution and encoding of new spatial information can be mitigated by having an initial activity state that is relatively close, in the high-dimensional space of such representations, to a sufficiently distinct and informative new spatial map. The alternative, an initial state characterized by sparse random activity, would require a much higher degree of reorganization through experience to achieve an operational map that is also statistically independent of previously learned representations.

Our model demonstrates that a readout constrained by network dynamics can achieve robust global remapping, for realignment parameters within the observed ranges, despite varying degrees of coherence. Interestingly, while the globally coherent realignment conditions did not orthogonalize spatial codes according to our pair-wise remapping measures (s1 and rs1 in Figure 4.12), they still produced substantial remapping. Coherent shift realignment (s1) yielded about 85% place and 80% and rate remapping; measured as population rate decorrelation, this yielded complete orthogonalization (Figure 4.9), though that is the result of a simple coordinate shift larger than the diameter of most place fields. Somewhat counterintuitively, when combined with rotational realignment (rs1) the pair-wise remapping measures decreased. This suggests that the rotation recovered some of the pairwise structure of the previous spatial code. This is not the case, however, in the modular conditions, where just two shifting modules produces around 95% place and 85% and rate remapping and adding rotation enhances overall remapping (Section 4.3.6). The efficacy of modular realignment depends on the connectivity patterns between modules and place units. Here, we create modules randomly so that a given place unit has inputs uniformly sampled from the available modules.

Nonetheless, even with significant amounts of coherence, our model is capable of amplifying small detailed changes to produce significant remapping.

Beyond pair-wise or rate-based measures as benchmarks of remapping, the subsets of active units that support the spatial code in distinct environments should be independently sampled from the place unit population. Both global remapping and task switching are accompanied by the recruitment of independent cell assemblies in CA3 as the neuronal basis of new spatial maps (Leutgeb et al., 2004, 2005; Jackson and Redish, 2007). Further, there is subregional differentiation, as CA1 and CA3 ensembles exhibit graded and discontinuous changes, respectively, in response to different degrees of environmental modification (Vazdarjanova and Guzowski, 2004). Remapping in CA1, then, is similar to the partial remapping that we observed with coherent shift realignment in which a small subset of place units randomly recode. However, we found that the same MEC realignment conditions that effectively orthogonalized spatial codes tended to do so by independent recruitment of active subsets of place units (Figure 4.10). This indicates that modular non-coherence successfully creates new spatial codes by enabling a new random subset of place units to develop active place fields.

If we assume, as some evidence suggests, that entorhinal realignment is in fact globally coherent, then we can consider some relevant implications of our results. First, we showed that competitive network dynamics can significantly disrupt the internal structure of a previous spatial map. So, cortical coherence does not necessarily entail redundant and ineffective remapping. Second, depending on either interpretation from Section 3.5.3, this mechanism is a single component of DG/CA3 processing of perforant path inputs. There are other mechanisms for pattern separation in the system that can certainly contribute to and enhance remapping (Acsady and Kali, 2007; Leutgeb et al., 2007; McHugh et al., 2007). For example, this

direct-input mechanism for encoding may be complemented by sparse inputs from strong mossy-fiber projections (cf. Treves and Rolls, 1992) as dentate granule cells become spatially selective with experience (Leutgeb et al., 2007). Third, remapping deficiencies are partly a result of considering only A–B remapping experiments. In actuality, animals visit numerous novel environments throughout their lifetimes. If we consider three environments (A, B, and C) explored successively, then the first-order comparisons A–B and B–C will share some residual structural similarity. However, transitively, the second-order A–C pair would be much less similar than the first-order comparisons. That is, some amount of redundancy between sequentially visited environments translates to effective orthogonalization of the representations for non-contiguously experienced environments.

4.4.3 Implications for episodic memory

Intriguingly, and somewhat speculatively, the spatial redundancy observed under coherent realignment could be framed as a feature of episodic encoding. Consider our three successively visited environments as the respective spatiotemporal contexts of several sequential episodes. The contiguous episodes, corresponding to environment pairs A–B and B–C, occured closer in time than non-contiguous episodes. Correspondingly, the pair-wise structural similarity of the spatial representations decreases with temporal distance between events. Various models of temporal context traces have been posited as a means of binding sequences of episodes through time (e.g., Howard and Kahana, 2002). Similarly, this redundancy for sequentially experienced environments could contribute a spatial component to spatiotemporal context during the encoding of new memories.

Spatial activity in hippocampus has long been theorized to have a critical role in the translation of behavior and experience into long-term episodic memory (Marr,

1971; Teyler and DiScenna, 1986; Aggleton and Brown, 1999) and experimental data have generally been supportive of such a role (Eichenbaum, 1999; Leutgeb et al., 2005; Dragoi and Buzsáki, 2006; Knierim et al., 2006; Teyler and Rudy, 2007). Thus, the relative effectiveness of hippocampal remapping, and its analogs in humans (Ekstrom et al., 2003) or other mammals, could be an important factor in mnemonic function. Subregion CA3, specifically, may serve to create configural associations between object or item representations and spatial representations (Knierim et al., 2006). Non-spatial inputs from lateral entorhinal cortex may modulate the CA3 population rate code while its place code may provide the associated spatial context (Hargreaves et al., 2005). Indeed, in the presence of cue changes or other minor modifications of familiar environments, the CA3 positional code remains coherent with the salient spatial context (Lee et al., 2004; Leutgeb and Leutgeb, 2007). This contextual coherence is functionally a form of pattern completion. Such pattern completion may be foundational to the particular sort of neural computation performed by CA3 in familiar environments (Marr, 1971), but it depends on an encoding process that provides an independent neuronal basis of activity in distinct contexts. That is, the decoding process of pattern completion requires an encoding process of pattern separation (Knierim et al., 2006; Leutgeb and Leutgeb, 2007). Thus, a fast, dynamic and input-sensitive readout of MEC realignment may provide a basis for the rapid reconfiguration of spatial representations necessary for episodic memory encoding.

Chapter 5

Conclusion

In Chapter 1, we introduced the mnemonic framework for the investigations that we pursued in later chapters. We discussed the conventional view that parallel information processing streams within MTL and the parahippocampal areas are integrated and associated within hippocampus. This view provides the groundwork for understanding episodic memory formation, but other critical forms of memory may be separately subserved by discrete elements within those processing streams. Familiarity-based recognition and episodic memory operate at different levels of abstraction and with vastly different computational requirements, but each are individually critical to normal cognition in mammals. We reviewed familiarity processing in perirhinal cortex and spatial representation in hippocampus in light of these considerations.

In Chapter 2, we presented a minimal model of familiarity discrimination that we argued is isomorphic to the familiarity computation performed by primate perirhinal cortex. We trained the model on an empirical dataset of semantic similarity as a proxy for the semantic feature space integrated by perirhinal cortex from its neocortical afferents. We found that the high degree of clustering and feature correlation within the dataset reduced the effective storage capacity of the network for performing recognition tasks. However, those correlations revealed

a word-frequency-based dependency that reproduced a known recognition memory effect in humans. This word-frequency mirror effect was apparent using a frequency-based criterion shift for recognition performance; we argued that this supports dual-process involvement in familiarity discrimination.

In Chapter 3, we presented a competitive model of spatial representation using recurrent inhibition as a putative mechanism for the hard-wired network dynamics underlying the initial formation of spatial maps in hippocampus. We showed that the model can be broadly tuned to an excitatory–inhibitory balance state that produces operationally relevant "prototype" spatial maps based on the concurrent response structure of simulated MEC grid-cell responses. The competitive balance of the resulting spatial codes is heterogeneous across the environment, enhancing the input-sensitivity of the outputs. While the entire response is activity-dependent, we found that the place code is almost instantaneously available whereas the rate code was generally slower to converge. In hippocampus, this mechanism could function as a real-time readout, driven by direct entorhinal input from layer II along the perforant path projection. Under conditions of remapping, it would be able to respond rapidly and contiguously with MEC realignment.

In Chapter 4, we explored the ability of these prototype maps to remap based on cortical realignment. To quantify remapping, we considered the pair-wise relative structure of the place and rate codes; this proves to be a stricter measure of change than dot-product or correlation measures. We found that coherent realignment can provide substantial, but not complete, remapping. However, the basis of this remapping is the random recoding of a minor subset of place units, the size of which depends on the magnitude of displacement. This is inconsistent with the independent recruitment of cell ensembles in global remapping. We found that the rotational component of realignment interferes with remapping for coherent

realignment, but not for non-coherent realignment. Implementing modularity as a source of global non-coherence, we found that a small number of modules is sufficient to produce effective and robust global remapping. Modularity also reduces the dependence of remapping on the magnitude of realignment. Finally, the fast, dynamic mechanism presented here may serve to bootstrap the distinct neuronal index of activity in novel environments required for the rapid formation of detailed episodic memory.

These investigations generally serve to illustrate the importance of the large-scale structure of the feature spaces that are integrated within models of high-order processing. In the computational biosciences, emphasis usually rests on complex and biologically realistic interactions at the expense of understanding or implementing the detailed structure of the inputs to the system. Throughout this work, our model design was guided by the converse notion, that relatively simple interactions may reveal insights hidden in complicated and structured input spaces. This approach places a particular burden on the interpretational framework in which the model and its results are couched, as the biology is always deeper and more complex than the model represents. However, this is a burden shared generally by all modeling approaches, regardless of apparent biological realism or plausibility. By shifting the emphasis of computational function to the input space, we can achieve a more integrative perspective of the model system and, perhaps, further insight into the computations that it performs.

Bibliography

Acsady, L. and S. Kali (2007). Models, structure, function: The transformation of cortical signals in the dentate gyrus. *Prog Brain Res 163*, 577–599.

Acsady, L., A. Kamondi, A. Sik, T. Freund, and G. Buzsáki (1998). GABAergic cells are the major postsynaptic targets of mossy fibers in the rat hippocampus. *J Neurosci 18*(9), 3386–3403.

Aggleton, J. and M. W. Brown (1999). Episodic memory, amnesia and the hippocampal-anterior thalamic axis. *Behav Brain Sci 22*(3), 425–+.

Aggleton, J. P. and M. W. Brown (2006). Interleaving brain systems for episodic and recognition memory. *Trends Cogn Sci 10*(10), 455–463.

Agnihotri, N. T., R. D. Hawkins, E. R. Kandel, and C. Kentros (2004). The long-term stability of new hippocampal place fields requires new protein synthesis. *Proc Nat Acad Sci USA 101*(10), 3656–3661.

Amaral, D. G., N. Ishizuka, and B. Claiborne (1990). Neurons, numbers and the hippocampal network. *Prog Brain Res 83*, 1–11.

Amit, D. J. (1989). *Modeling Brain Function*. Cambridge, U. K.: Cambridge University Press.

Arndt, J. and L. M. Reder (2002). Word frequency and receiver operating characteristic curves in recognition memory: Evidence for a dual-process interpretation. *J Exp Psychol Learn Mem Cogn 28*(5), 830–42.

Barry, C., R. Hayman, N. Burgess, and K. J. Jeffery (2007). Experience-dependent rescaling of entorhinal grids. *Nat Neurosci 10*(6), 682–684.

Benjamin, A. S. and S. Bawa (2004). Distractor plausibility and criterion placement in recognition. *J Mem Lang 51*, 159–172.

Bogacz, R. and M. W. Brown (2002). Capacity of perirhinal cortex network for recognising frequently repeating stimuli. *Neurocomputing 44–46*, 337–342.

Bogacz, R. and M. W. Brown (2003). Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus 13*(4), 494–524.

Bogacz, R., M. W. Brown, and C. Giraud-Carrier (2001a). A familiarity discrimination algorithm inspired by computations of the perirhinal cortex. *Lect Notes Comput Sc 2036*, 428–441.

Bogacz, R., M. W. Brown, and C. Giraud-Carrier (2001b). Model of familiarity discrimination in perirhinal cortex. *J Comput Neurosci 10*(1), 5–23.

Bostock, E., R. U. Muller, and J. L. Kubie (1991). Experience-dependent modifications of hippocampal place cell firing. *Hippocampus 1*(2), 193–205.

Brown, M. W. and Z. I. Bashir (2002). Evidence concerning how neurons of the perirhinal cortex may effect familiarity discrimination. *Phil Trans R Soc Lond B 357*, 1083–1095.

Brown, M. W. and J.-Z. Xiang (1998). Recognition memory: Neuronal substrates of the judgement of prior occurrence. *Prog Neurobiol 55*(2), 149–189.

Brun, V. H., S. Leutgeb, H.-Q. Wu, R. Schwarcz, M. P. Witter, E. I. Moser, and M.-B. Moser (2008). Impaired spatial representation in CA1 after lesion of direct input from entorhinal cortex. *Neuron 57*(2), 290–302.

Buckmaster, P. S. and P. A. Schwartzkroin (1995). Interneurons and inhibition in the dentate gyrus of the rat in vivo. *J Neurosci 15*(1), 774–789.

Burgess, N., C. Barry, and J. O'Keefe (2007). An oscillatory interference model of grid cell firing. *Hippocampus 17*(9), 801–812.

Buzsáki, G., K. Kaila, and M. Raichle (2007). Inhibition and brain work. *Neuron 56*(5), 771–783.

Canto, C. B., F. G. Wouterlood, and M. P. Witter (2008). What does the anatomical organization of the entorhinal cortex tell us? *Neural Plasticity 2008*, 381243.

Carey, B. (2008). H. M., an Unforgettable Amnesiac, Dies at 82. *The New York Times Dec 4, 2008*, A1.

Cho, K., N. Kemp, J. Noel, J. P. Aggleton, M. W. Brown, and Z. I. Bashir (2000). A new form of long-term depression in the perirhinal cortex. *Nat Neurosci 3*(2), 150–156.

Colgin, L. L., E. I. Moser, and M.-B. Moser (2008). Understanding memory through hippocampal remapping. *Trends Neurosci 31*(9), 469–477.

Cressant, A., R. U. Muller, and B. Poucet (2002). Remapping of place cell firing patterns after maze rotations. *Exp Brain Res 143*(4), 470–479.

Croll, S. D., P. E. Sharp, and E. Bostock (1992). Evidence for NMDA receptor involvement in environmentally induced dentate gyrus plasticity. *Hippocampus 2*(1), 23–28.

de Zubicaray, G. I., K. L. McMahon, M. M. Eastburn, S. Finnigan, and M. S. Humphreys (2005). fMRI evidence of word frequency and strength effects in recognition memory. *Brain Res Cogn Brain Res 24*(3), 587–598.

Dolorfo, C. L. and D. G. Amaral (1998). Entorhinal cortex of the rat: Topographic organization of the cells of origin of the perforant path projection to the dentate gyrus. *J Comp Neurol 398*(1), 25–48.

Dragoi, G. and G. Buzsáki (2006). Temporal encoding of place sequences by hippocampal cell assemblies. *Neuron 50*(1), 145–157.

Eichenbaum, H. (1999). The hippocampus and mechanisms of declarative memory. *Behav Brain Res 103*(2), 123–133.

Ekstrom, A. D., M. J. Kahana, J. B. Caplan, T. A. Fields, E. A. Isham, E. L. Newman, and I. Fried (2003). Cellular networks underlying human spatial navigation. *Nature 425*(6954), 184–188.

Erickson, C. A., B. Jagadeesh, and R. Desimone (2000). Clustering of perirhinal neurons with similar properties following visual experience in adult monkeys. *Nature Neurosci 3*(11), 1143–1148.

Fahy, F., I. Riches, and M. W. Brown (1993). Neuronal activity related to visual recognition memory: Long-term memory and the encoding of recency and familiarity information in the primate anterior and medial inferior temporal and rhinal cortex. *Exp Brain Res 96*, 457–472.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Frank, L. M., G. B. Stanley, and E. N. Brown (2004). Hippocampal plasticity across multiple days of exposure to novel environments. *J Neurosci 24*(35), 7681–7689.

Franzius, M., R. Vollgraf, and L. Wiskott (2006). From grids to places. *J Comput Neurosci 22*, 297–299.

Fyhn, M., T. Hafting, A. Treves, E. I. Moser, and M.-B. Moser (2007). Hippocampal remapping and grid realignment in entorhinal cortex. *Nature 446*(7132), 190–194.

Fyhn, M., S. Molden, M. P. Witter, E. I. Moser, and M.-B. Moser (2004). Spatial representation in the entorhinal cortex. *Science 305*(5688), 1258–1264.

Gaffan, D. (1994). Dissociated effects of perirhinal cortex ablation, fornix transection and amygdalectomy – evidence for multiple memory-systems in the primate temporal-lobe. *Exp Brain Res 99*(3), 411–422.

Gerstner, W. and L. F. Abbott (1997). Learning navigational maps through potentiation and modulation of hippocampal place cells. *J Comput Neurosci 4*, 79–94.

Giocomo, L. M. and M. E. Hasselmo (2008). Computation by oscillations: Implications of experimental data for theoretical models of grid cells. *Hippocampus 18*(12), 1186–1199.

Giocomo, L. M., E. A. Zilli, E. Fransén, and M. E. Hasselmo (2007). Temporal frequency of subthreshold oscillations scales with entorhinal grid cell field spacing. *Science 315*(5819), 1719–1722.

Glanzer, M. and J. K. Adams (1985). The mirror effect in recognition memory. *Mem Cognit 13*(1), 8–20.

Glanzer, M. and J. K. Adams (1990). The mirror effect in recognition memory: Data and theory. *J Exp Psychol Learn Mem Cogn 16*(1), 5–16.

Glanzer, M., J. K. Adams, G. J. Iverson, and K. Kisok (1993). The regularities of recognition memory. *Psychol Rev 100*(3), 546–567.

Guttentag, R. E. and D. Carroll (1994). Identifying the basis for the word-frequency effect in recognition memory. *Memory 2*(3), 255–273.

Guttentag, R. E. and D. Carroll (1997). Recollection-based recognition: Word frequency effects. *J Mem Lang 37*(4), 502–516.

Haberman, R. P., H. J. Lee, C. Colantuoni, M. T. Koh, and M. Gallagher (2008). Rapid encoding of new information alters the profile of plasticity-related mRNA transcripts in the hippocampal CA3 region. *Proc Nat Acad Sci USA 105*(30), 10601–10606.

Hafting, T., M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature 436*(7052), 801–806.

Hafting, T., M. Fyhn, T. Solstad, M.-B. Moser, and E. I. Moser (2008). Coherence in ensembles of entorhinal grid cells across hemispheres. *Soc Neurosci Abstract 94.7*.

Han, J.-H., S. A. Kushner, A. P. Yiu, C. J. Cole, A. Matynia, R. A. Brown, R. L. Neve, J. F. Guzowski, A. J. Silva, and S. A. Josselyn (2007). Neuronal competition and selection during memory formation. *Science 316*(5823), 457–460.

Hargreaves, E. L., G. Rao, I. Lee, and J. J. Knierim (2005). Major dissociation between medial and lateral entorhinal input to dorsal hippocampus. *Science 308*(5729), 1792–1794.

Hasselmo, M. E. (1995). Neuromodulation and cortical function: Modeling the physiological basis of behavior. *Behav Brain Res 67*(1–27).

Hasselmo, M. E., L. M. Giocomo, and E. A. Zilli (2007). Grid cell firing may arise from interference of theta frequency membrane potential oscillations in single neurons. *Hippocampus 17*(12), 1252–1271.

Hasselmo, M. E. and E. Schnell (1994). Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: Computational modeling and brain slice physiology. *J Neurosci 14*(6), 3898–3914.

Hayman, R. M. and K. J. Jeffery (2008). How heterogeneous place cell responding arises from homogeneous grids: A contextual gating hypothesis. *Hippocampus 18*(12), 1301–1313.

Heit, E., N. Brockdorff, and K. Lamberts (2003). Adaptive changes of response criterion in recognition memory. *Psychon Bull Rev 10*(3), 718–723.

Hetherington, P. A. and M. L. Shapiro (1997). Hippocampal place fields are altered by the removal of single visual cues in a distance-dependent manner. *Behav Neurosci 111*(1), 20–34.

Hill, A. J. (1978). First occurrence of hippocampal spatial firing in a new environment. *Exp Neurol 62*(2), 282–297.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *J Exp Psychol Learn Mem Cogn 21*(2), 302–313.

Hirshman, E., J. Fisher, T. Henthorn, J. Arndt, and A. Passannante (2002). Midazolam amnesia and dual-process models of the word-frequency mirror effect. *J Mem Lang 47*, 499–516.

Hodges, J. R. and K. Patterson (2007). Semantic dementia: A unique clinicopathological syndrome. *Lancet Neurol 6*(11), 1004–1014.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc Nat Acad Sci USA 84*, 8429–8433.

Howard, M. W. and M. J. Kahana (2002). A distributed representation of temporal context. *J Math Psychol 46*(3), 269–299.

Humphreys, M. S., R. Pike, J. D. Bain, and G. Tehan (1989). Global matching: A comparison of the SAM, Minerva II, Matrix, and TODAM models. *J Math Psychol 33*, 36–67.

Hwang, G. M., J. Jacobs, A. Geller, J. Danker, R. Sekuler, and M. J. Kahana (2005). EEG correlates of verbal and nonverbal stimuli in working memory. *Behav Brain Funct 1*(1), 20 [Epub].

Insausti, R. and D. G. Amaral (2008). Entorhinal cortex of the monkey: IV. Topographical and laminar organization of cortical afferents. *J Comp Neurol 509*(6), 608–641.

Jackson, J. and A. D. Redish (2007). Network dynamics of hippocampal cell-assemblies resemble multiple spatial maps within single tasks. *Hippocampus 17*(12), 1209–1229.

Jaffe, D. B. and R. Gutierrez (2007). Mossy fiber synaptic transmission: Communication from the dentate gyrus to area CA3. *Prog Brain Res 163*, 109–132.

Jezek, K., A. Treves, M.-B. Moser, and E. I. Moser (2008). Abrupt and stuttering transitions between hippocampal representations. *Soc Neurosci Abstract 94.11*.

Kahana, M. J., D. S. Rizzuto, and A. Schneider (2005). An analysis of the recognition-recall relation in four distributed memory models. *J Exp Psychol Learn Mem Cogn 31*, 933–953.

Kahana, M. J. and R. Sekuler (2002). Recognizing spatial patterns: A noisy exemplar approach. *Vision Res 42*(18), 2177–2192.

Karlsen, P. and J. Snodgrass (2004). The word-frequency paradox for recall/recognition occurs for pictures. *Psychol Res 68*(4), 271–276.

Karlsson, M. and L. Frank (2008). Network dynamics underlying the formation of sparse, informative representations in the hippocampus. *J Neurosci 28*(52), 14271–14281.

Kentros, C., E. Hargreaves, R. D. Hawkins, E. R. Kandel, M. Shapiro, and R. V. Muller (1998). Abolition of long-term stability of new hippocampal place cell maps by NMDA receptor blockade. *Science 280*(5372), 2121–2126.

Kesner, R. P., P. E. Gilbert, and G. V. Wallenstein (2000). Testing neural models of memory with behavioral experiments. *Curr Opin Neurobiol 10*, 260–265.

Kiani, R., H. Esteky, K. Mirpour, and K. Tanaka (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J Neurophysiol 97*(6), 4296–4309.

Kjelstrup, K. B., T. Solstad, V. H. Brun, T. Hafting, S. Leutgeb, M. P. Witter, E. I. Moser, and M.-B. Moser (2008). Finite scale of spatial representation in the hippocampus. *Science 321*(5885), 140–143.

Knierim, J. J., H. S. Kudrimoti, and B. L. McNaughton (1998). Interactions between idiothetic cues and external landmarks in the control of place cells and head direction cells. *J Neurophysiol 80*(1), 425–446.

Knierim, J. J., I. Lee, and E. L. Hargreaves (2006). Hippocampal place cells: Parallel input streams, subregional processing, and implications for episodic memory. *Hippocampus 16*(9), 755–764.

Knierim, J. J. and B. L. McNaughton (2001). Hippocampal place-cell firing during movement in three-dimensional space. *J Neurophysiol 85*(1), 105–116.

Kubie, J. L., R. U. Muller, and E. Bostock (1990). Spatial firing properties of hippocampal theta cells. *J Neurosci 10*(4), 1110–1123.

Kučera, H. and W. N. Francis (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Lacroix, J. P. W., J. M. J. Murre, E. O. Postma, and H. J. van den Herik (2006). Modeling recognition memory using the similarity structure of natural input. *Cogn Sci 30*, 121–145.

Lee, I., M. R. Hunsaker, and R. P. Kesner (2005). The role of hippocampal subregions in detecting spatial novelty. *Behav Neurosci 119*(1), 145–153.

Lee, I. and R. P. Kesner (2002). Differential contribution of NMDA receptors in hippocampal subregions to spatial working memory. *Nat Neurosci 5*(2), 162–168.

Lee, I. and J. J. Knierim (2007). The relationship between the field-shifting phenomenon and representational coherence of place cells in CA1 and CA3 in a cue-altered environment. *Learn Mem 14*(11), 807–815.

Lee, I., G. Rao, and J. J. Knierim (2004). A double dissociation between hippocampal subfields: Differential time course of CA3 and CA1 place cells for processing changed environments. *Neuron 42*(5), 803–815.

Lee, I., D. Yoganarasimha, G. Rao, and J. J. Knierim (2004). Comparison of population coherence of place cells in hippocampal subfields CA1 and CA3. *Nature 430*(6998), 456–459.

Leutgeb, J. K., S. Leutgeb, M.-B. Moser, and E. I. Moser (2007). Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science 315*(5814), 961–966.

Leutgeb, J. K., S. Leutgeb, A. Tashiro, E. I. Moser, and M.-B. Moser (2007). The encoding of novelty in the dentate gyrus and CA3 network. *Soc Neurosci Abstract 93.9*.

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

Leutgeb, S. and J. K. Leutgeb (2007). Pattern separation, pattern completion, and new neuronal codes within a continuous CA3 map. *Learn Mem 14*, 745–757.

Leutgeb, S., J. K. Leutgeb, C. A. Barnes, E. I. Moser, B. L. McNaughton, and M.-B. Moser (2005). Independent codes for spatial and episodic memory in hippocampal neuronal ensembles. *Science 309*(5734), 619–623.

Leutgeb, S., J. K. Leutgeb, E. I. Moser, and M.-B. Moser (2006). Fast rate coding in hippocampal CA3 cell ensembles. *Hippocampus 16*(9), 765–774.

Leutgeb, S., J. K. Leutgeb, A. Treves, M.-B. Moser, and E. I. Moser (2004). Distinct ensemble codes in hippocampal areas CA3 and CA1. *Science 305*(5688), 1295–1298.

Levy, D. A., P. J. Bayley, and L. R. Squire (2004). The anatomy of semantic knowledge: Medial vs. lateral temporal lobe. *Proc Nat Acad Sci USA 101*(17), 6710–6715.

Levy, W. B. (1989). A computational approach to hippocampal function. In R. D. Hawkins and G. H. Bower (Eds.), *Computational models of learning in simple neural systems*, pp. 243–305. Academic Press: New York, NY.

Levy, W. B., C. M. Colbert, and N. L. Desmond (1995). Another network model bites the dust: Entorhinal inputs are no more than weakly excitatory in the hippocampal CA1 region. *Hippocampus 5*(2), 137–140.

Li, L., E. K. Miller, and R. Desimone (1993). The representation of stimulus familiarity in anterior inferior temporal cortex. *J Neurophysiol 68*(6), 1918–1929.

Li, S., W. K. Cullen, R. Anwyl, and M. J. Rowan (2003). Dopamine-dependent facilitation of LTP induction in hippocampal CA1 by exposure to spatial novelty. *Nat Neurosci 6*(5), 526–531.

Malmberg, K. J. and T. O. Nelson (2003). The word frequency effect for recognition memory and the elevated-attention hypothesis. *Mem Cognit 31*, 35–43.

Malmberg, K. J., M. Steyvers, J. D. Stephens, and R. M. Shiffrin (2002). Feature frequency effects in recognition memory. *Mem Cognit 30*(4), 607–613.

Malmberg, K. J., R. Zeelenberg, and R. M. Shiffrin (2004). Turning up the noise or turning down the volume? On the nature of the impairment of episodic recognition memory by midazolam. *J Exp Psychol Learn Mem Cogn 30*(2), 540–549.

Markus, E. J., Y. L. Qin, B. Leonard, W. E. Skaggs, B. L. McNaughton, and C. A. Barnes (1995). Interactions between location and task affect the spatial and directional firing of hippocampal neurons. *J Neurosci 15*(11), 7079–7094.

Marr, D. (1971). Simple memory: A theory for archicortex. *Philos Trans R Soc Lond B Biol Sci 262*(841), 23–81.

Marshall, L., D. A. Henze, H. Hirase, X. Leinekugel, G. Dragoi, and G. Buzsáki (2002). Hippocampal pyramidal cell–interneuron spike transmission is frequency dependent and responsible for place modulation of interneuron discharge. *J Neurosci 22*(2), RC197.

Maurer, A. P., S. L. Cowen, S. N. Burke, C. A. Barnes, and B. L. McNaughton (2006). Phase precession in hippocampal interneurons showing strong functional coupling to individual pyramidal cells. *J Neurosci 26*(52), 13485–13492.

McClelland, J. L. and M. Chappell (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychol Rev 1-5*(4), 724–760.

McHugh, T. J., M. W. Jones, J. J. Quinn, N. Balthasar, R. Coppari, J. K. Elmquist, B. B. Lowell, M. S. Fanselow, M. A. Wilson, and S. Tonegawa (2007). Dentate gyrus NMDA receptors mediate rapid pattern separation in the hippocampal network. *Science 317*(5834), 94–99.

McNaughton, B. L., F. P. Battaglia, O. Jensen, E. I. Moser, and M.-B. Moser (2006). Path integration and the neural basis of the 'cognitive map'. *Nat Rev Neurosci 7*, 663–678.

Mehta, M. R., C. A. Barnes, and B. L. McNaughton (1997). Experience-dependent, asymmetric expansion of hippocampal place fields. *Proc Natl Acad Sci U S A 94*(16), 8918–8921.

Mehta, M. R., M. C. Quirk, and M. A. Wilson (2000). Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron 25*(3), 707–715.

Miller, E. K. (2000). Organization through experience. *Nature Neurosci 3*(11), 1066–1068.

Miller, E. K., L. Li, and R. Desimone (1991). A neural mechanism for working and recognition memory in inferior temporal cortex. *Science 254*(5036), 1377–1379.

Miller, E. K., A. Nieder, D. J. Freedman, and J. D. Wallis (2003). Neural correlates of categories and concepts. *Curr Opin Neurobiol 13*, 198–203.

Miller, M. B. and G. L. Wolford (1999). Theoretical commentary: The role of criterion shift in false memory. *Psychol Rev 106*(2), 398–405.

Mintzner, M. Z. (2003). Triazolam-induced amnesia and the word-frequency effect in recognition memory: Support for a dual process account. *J Mem Lang 48*, 596–602.

Moita, M. A. P., S. Rosis, Y. Zhou, J. E. LeDoux, and H. T. Blair (2004). Putting fear in its place: Remapping of hippocampal place cells during fear conditioning. *J Neurosci 24*(31), 7015–7023.

Molter, C. and Y. Yamaguchi (2008). Entorhinal theta phase precession sculpts dentate gyrus place fields. *Hippocampus 18*(9), 919–930.

Monaco, J. D., L. F. Abbott, and M. J. Kahana (2007). Lexico-semantic structure and the word-frequency effect in recognition memory. *Learn Mem 14*(3), 204–213.

Monaco, J. D. and W. B. Levy (2003). T-maze training of a recurrent CA3 model reveals the necessity of novelty-based modulation of LTP in hippocampal region CA3. In *Proceedings of IJCNN*, Portland, OR, pp. 1655–1660. IEEE.

Moser, E. I., E. Kropff, and M.-B. Moser (2008). Place cells, grid cells, and the brain's spatial representation system. *Annu Rev Neurosci 31*(1), 69–89.

Muller, R. U. and J. L. Kubie (1987). The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *J Neurosci 7*(7), 1951–1968.

Muller, R. U., J. L. Kubie, and J. B. J. Ranck (1987). Spatial firing patterns of hippocampal complex-spike cells in a fixed environment. *J Neurosci 7*(7), 1935–1950.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychol Rev 89*, 609–626.

Murdock, B. B. (1998). The mirror effect and attention-likelihood theory: A reflective analysis. *J Exp Psychol Learn Mem Cogn 24*(2), 524–534.

Murray, E. A. and T. J. Bussey (1999). Perceptual-mnemonic functions of the perirhinal cortex. *Trends Cogn Sci 3*(4), 142–151.

Murray, E. A., T. J. Bussey, and L. M. Saksida (2007). Visual perception and memory: A new view of medial temporal lobe function in primates and rodents. *Annu Rev Neurosci 30*, 99–122.

Nakazawa, K., L. D. Sun, M. C. Quirk, L. Rondi-Reig, M. A. Wilson, and S. Tonegawa (2003). Hippocampal CA3 NMDA receptors are crucial for memory acquisition of one-time experience. *Neuron 38*(2), 305–315.

Nelson, D., C. McEvoy, and T. Schreiber (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behav Res Meth Instr 36*(3), 402–407.

Nitz, D. and B. McNaughton (2004). Differential modulation of CA1 and dentate gyrus interneurons during exploration of novel environments. *J Neurophysiol 91*(2), 863–872.

Norman, K. A., E. L. Newman, and A. J. Perotte (2005). Methods for reducing interference in the complementary learning systems model: Oscillating inhibition and autonomous memory rehearsal. *Neural Networks 18*(9), 1212–1228.

Norman, K. A. and R. C. O'Reilly (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychol Rev 110*(4), 611–646.

O'Keefe, J. (1976). Place units in the hippocampus of the freely moving rat. *Exp Neurol 51*(1), 78–109.

O'Keefe, J. and N. Burgess (2005). Dual phase and rate coding in hippocampal place cells: Theoretical significance and relationship to entorhinal grid cells. *Hippocampus 15*, 853–866.

O'Keefe, J. and J. Dostrovsky (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Res 34*(1), 171–175.

O'Keefe, J. and L. Nadel (1978). *The Hippocampus as a Cognitive Graph*. Oxford, UK: Clarendon Press.

O'Reilly, R. C. and J. L. McClelland (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus 4*(6), 661–682.

Paulsen, O. and E. I. Moser (1998). A model of hippocampal memory encoding and retrieval: GABAergic control of synaptic plasticity. *Trends Neurosci 21*(7), 273–278.

Paz-Villagran, V., E. Save, and B. Poucet (2004). Independent coding of connected environments by place cells. *Eur J Neurosci 20*(5), 1379–1390.

Pelletier, J. G. and J.-C. Lacaille (2008). Long-term synaptic plasticity in hippocampal feedback inhibitory networks. *Prog Brain Res 169*, 241–250.

Quirk, G. J., R. U. Muller, and J. L. Kubie (1990). The firing of hippocampal place cells in the dark depends on the rat's recent experience. *J Neurosci 10*(6), 2008–2017.

Ramón y Cajal, S. (1901). *Studien uber die Hirnrinde des Menschen*. Leipzig: J. A. Barth.

Ranck, J. B. J. (1973). Studies on single neurons in dorsal hippocampal formation and septum in unrestrained rats. I. Behavioral correlates and firing repertoires. *Exp Neurol 41*(2), 461–531.

Reder, L. M., A. Nhoujvanisvong, C. D. Schunn, M. S. Ayers, P. Angstadt, and K. Hiraki (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. *J Exp Psychol Learn Mem Cogn 26*(2), 294–320.

Roediger, H. L. and K. B. McDermott (1999). False alarms about false memories. *Psychol Rev 106*(2), 406–410.

Rolls, E. T., S. M. Stringer, and T. Elliot (2006). Entorhinal grid cells can map to hippocampal place cells by competitive learning. *Network-Comp Neural 17*(4), 447–465.

Rugg, M. D. and A. P. Yonelinas (2003). Human recognition memory: A cognitive neuroscience perspective. *Trends Cogn Sci 7*(7), 313–319.

Savelli, F. and J. J. Knierim (2007). A Hebbian model of the formation of place fields from MEC grid inputs. *Soc Neurosci Abstract 205.21*.

Schmolck, H., E. A. Kensinger, S. Corkin, and L. R. Squire (2002). Semantic knowledge in patient H.M. and other patients with bilateral medial and lateral temporal lobe lesions. *Hippocampus 12*(4), 520–533.

Schulman, A. I. (1967). Word length and rarity in recognition memory. *Psychon Soc 9*, 211–212.

Schwartz, G., M. W. Howard, B. Jing, and M. J. Kahana (2005). Shadows of the past: Temporal retrieval effects in recognition memory. *Psychol Sci 16*(11), 898–904.

Scoville, W. B. and B. Milner (1957). Loss of recent memory after bilateral hippocampal lesions. *J Neurol Neurosurg Psychiatry 20*(1), 11–21.

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *J Verb Learn Verb Be 6*(1), 156–+.

Shiffrin, R. M., R. Ratcliff, and S. Clark (1990). The list-strength effect: II. Theoretical mechanisms. *J Exp Psychol Learn Mem Cogn 16*, 179–195.

Shiffrin, R. M. and M. Steyvers (1997). A model for recognition memory: REM - retrieving effectively from memory. *Psychon Bull Rev 4*(2), 145–166.

Sik, A., M. Penttonen, and G. Buzsáki (1997). Interneurons in the hippocampal dentate gyrus: An in vivo intracellular study. *Eur J Neurosci 9*(3), 573–588.

Sobotka, S. and J. L. Ringo (1993). Investigation of long term recognition and association memory in unit responsses from inferotemporal cortex. *Exp Brain Res 96*, 28–38.

Sobotka, S. and J. L. Ringo (1996). Mnemonic responses of single units recorded from monkey inferotemporal cortex, accessed via transcommissural versus direct pathways: A dissociation between unit activity and behavior. *J Neurosci 16*(13), 4222–4230.

Sohal, V. S. and M. E. Hasselmo (2000). A model for experience-dependent changes in the responses of inferotemporal neurons. *Network Comp Neural 11*(3), 169–190.

Solstad, T., E. I. Moser, and G. T. Einevoll (2006). From grid cells to place cells: A mathematical model. *Hippocampus 16*(12), 1026–1031.

Squire, L. R., C. E. L. Stark, and R. E. Clark (2004). The medial temporal lobe. *Annu Rev Neurosci 27*, 279–306.

Squire, L. R., J. T. Wixted, and R. E. Clark (2007). Recognition memory and the medial temporal lobe: a new perspective. *Nat Rev Neurosci 8*(11), 872–883.

Squire, L. R. and S. M. Zola (1998). Episodic memory, semantic memory, and amnesia. *Hippocampus 8*(3), 205–211.

Standing, L. (1973). Learning 10,000 pictures. *Q J Exp Psychol 25*, 207–222.

Steyvers, M. and K. J. Malmberg (2003). The effect of normative context variability on recognition memory. *J Exp Psychol Learn Mem Cogn 29*(5), 760–766.

Steyvers, M., R. M. Shiffrin, and D. L. Nelson (2004). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Cognitive Psychology and its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer.* Washington, DC: American Psychological Association.

Steyvers, M. and J. B. Tenenbaum (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cogn Sci 29*, 41–78.

Strack, F., J. Förster, and L. Werth (2005). "Know thyself!" The role of idiosyncratic self-knowledge in recognition memory. *J Mem Lang 52*, 628–638.

Stretch, V. and J. T. Wixted (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *J Exp Psychol Learn Mem Cogn 24*(6), 1379–1396.

Suzuki, W. A. (1996). The anatomy, physiology and functions of the perirhinal cortex. *Curr Opin Neurobiol 6*(2), 179–186.

Teyler, T. J. and P. DiScenna (1986). The hippocampal memory indexing theory. *Behav Neurosci 100*(2), 147–154.

Teyler, T. J. and J. W. Rudy (2007). The hippocampal indexing theory and episodic memory: Updating the index. *Hippocampus 17*(12), 1158–1169.

Treves, A. and E. T. Rolls (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus 2*(2), 189–199.

van Strien, N. M., N. L. M. Cappaert, and M. P. Witter (2009). The anatomy of memory: An interactive overview of the parahippocampal-hippocampal network. *Nat Rev Neurosci 10*(4), 272–282.

Vazdarjanova, A. and J. F. Guzowski (2004). Differences in hippocampal neuronal population responses to modifications of an environmental context: Evidence for distinct, yet complementary, functions of CA3 and CA1 ensembles. *J Neurosci 24*(29), 6489–6496.

Wallenstein, G. V. and M. E. Hasselmo (1997). GABAergic modulation of hippocampal population activity: Sequence learning, place field development, and the phase precession effect. *J Neurophysiol 78*(1), 393–408.

Walling, S. G., K. Bromley, and C. W. Harley (2006). Glycogen phosphorylase reactivity in the entorhinal complex in familiar and novel environments: Evidence for labile glycogenolytic modules in the rat. *J Chem Neuroanat 31*(2), 108–113.

Watts, D. J. and S. H. Strogatz (1998). Collective dynamics of 'small-world' networks. *Nature 393*(6684), 440–442.

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. New York: Oxford University Press.

Wickens, T. D. and E. Hirshman (2000). False memories and statistical decision theory: Comment on Miller and Wolford (1999) and Roediger and McDermott (1999). *Psychol Rev 107*(2), 377–383.

Wills, T. J., C. Lever, F. Cacucci, N. Burgess, and J. O'Keefe (2005). Attractor dynamics in the hippocampal representation of the local environment. *Science 308*(5723), 873 – 876.

Wilson, M. A. and B. L. McNaughton (1993). Dynamics of the hippocampal ensemble code for space. *Science 261*(5124), 1055–1058.

Witter, M. P. (2007). Intrinsic and extrinsic wiring of CA3: Indications for connectional heterogeneity. *Learn Mem 14*(705–713).

Witter, M. P. and E. I. Moser (2006). Spatial representation and the architecture of the entorhinal cortex. *Trends Neurosci 29*(12), 671–678.

Wixted, J. T. and V. Stretch (2000). The case against a criterion-shift account of false memory. *Psychol Rev 107*(2), 368–376.

Xiang, J.-Z. and M. W. Brown (1998). Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacology 37*, 657–676.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *J Mem Lang 46*, 441–517.

Zaki, S. and R. Nosofsky (2001). Exemplar accounts of blending and distinctiveness effects in perceptual old-new recognition. *J Exp Psychol Learn Mem Cogn 27*(4), 1022–1041.

Zinyuk, L., S. Kubik, Y. Kaminsky, A. A. Fenton, and J. Bures (2000). Understanding hippocampal activity by using purposeful behavior: Place navigation induces place cell discharge in both task-relevant and task-irrelevant spatial reference frames. *Proc Natl Acad Sci USA 97*(7), 3771–3776.